Music Thumbnailing by Hooks

Arianne N. van Nieuwenhuijsen
6291384

Master thesis
Credits: 45 ECTS

Master Artificial Intelligence

Utrecht University

Daily supervisor
dr. J.A. (Ashley) Burgoyne

Institute for Logic, Language & Computation (ILLC)
University of Amsterdam
Science Park 107
1098 XG Amsterdam

Project supervisor (first examinator)
dr. F. (Frans) Wiering

Second examinator
dr. A. (Anja) Volk

Utrecht University
Buys Ballotgebouw
3584 CC Utrecht

July 12, 2019

Abstract

Music or audio thumbnailing is the procedure of finding a continuous fragment that can represent the whole musical piece. This study proposes to create thumbnails based on the perception of listeners to identify the most memorable and distinguishable fragment. This aligns with the cognitive definition of hooks, the catchiest part of a song. This study tested whether audio features previously used to define catchiness collude with representativeness. First, a user study was carried out to assign a score for representativeness and familiarity to fragments. Thereafter, audio features derived with the CATCHY toolbox were used to approximate these scores. The results indicate that features measuring intensity, commonality, and recurrence influence representativeness positively. This matches previous results regarding catchiness. Additionally, familiarity did not seem to have an impact and no preferred segmentation method was found. Lastly, a new music thumbnailing method is proposed based on the features that could approximate representativeness the best.

Acknowledgement

This thesis would not have been possible without the great deal of support that I have received. First, I would like to thank Muziekweb for the opportunity to work together on an interesting musical subject, and the music data and support for the survey they have provided.

I also want to thank my supervisors Ashley Burgoyne and Frans Wiering for their guidance, insightful feedback and discussions throughout this project. Without them I would not have been able to explore this subject.

Lastly, I would like thank my second examiner Anja Volk for thoroughly reading the final versions and Music Cognition Group for the sharp discussions I had.

Contents

Chapter 1

Introduction

With the rise of the digital age, more and more music is becoming available. Streaming services and websites with music videos make music readily accessible to the public. However, with the availability of so much music, the need to navigate through it easily becomes more and more important. One approach to tackle this problem is to create music thumbnails. Music thumbnailing, or audio thumbnailing, is the procedure of finding a continuous segment within a musical piece which represents the whole piece (Chai and Vercoe, 2003; Cooper and Foote, 2002; Huang et al., 2017; Müller et al., 2013). By using these shorter fragments of audio, music thumbnails make it easier for users to navigate through loads of music without having to spend too much time on listening to complete musical pieces.

A tangible example of how music thumbnails can be used is displayed at Muziekweb,[1] a Dutch music library that aims to make information about music available for everyone. On their website, excerpts can be played to get a sense of the musical piece. To be able to assess the musical pieces, choosing well-fitting representative music thumbnails becomes a must. However, currently these excerpts are chosen randomly, which makes it likely that these excerpts do not represent the musical pieces very well.

This study aims to improve on the current Muziekweb thumbnailing method by creating music thumbnails via the detection of the most catchy segment of a musical piece. The inspiration to look into catchiness comes from the project Hooked,[2] which aimed to detect features defining catchiness in music (Burgoyne et al., 2013; Van Balen et al., 2015a). Suggestions have already been made on the potential of catchiness for music search engines (Honing, 2010) and previous research also

[1] https://www.muziekweb.nl/

[2] http://www.hookedonmusic.org.uk/

indicated that the best thumbnails could be those segments containing the most memorable and distinguishable part of the musical piece (Chai and Vercoe, 2003; Huang et al., 2017). This aligns with the cognitive definition of hooks as used in the Hooked project; hooks are the most salient segments in a musical piece, making them the most recognisable part of a song (Burgoyne et al., 2013). As the most catchiest part of a musical piece, the hook should be the most recognisable part and could therefore be a good representation of the musical piece for human listeners.

To use the notion of hooks and catchiness for audio thumbnailing, this study has a similar setup to the study by Müllensiefen and Halpern (2014) into what features and context facilitate the recognition of novel melodies. This means that first a user study is performed to obtain information on the representativity of different fragments of the same tunes. This user study is similar to one of the tasks in Hooked (Burgoyne et al., 2013). Thereafter, several features are extracted from the audio fragments via the CATCHY toolbox by Van Balen et al. (2015a), which has previously been used to analyse the Hooked data. To make the features easier to interpret and to explain the variance in the data with less features, dimensionality reduction is used to obtain factors. These factors are used to see whether they can approximate the score obtained via the user study in a linear model. Afterwards, the features that describe the representativity the best are used to create a function to rate the representativity of fragments of a song. With this scoring, the best-rated fragment can be detected and used as audio thumbnail.

Putting this study in context, it falls within the scope of Music Information Retrieval (MIR) and music cognition. MIR is an interdisciplinary research field that uses computational approaches to deal with music data in digital form (Burgoyne et al., 2016; Futrelle and Downie, 2002). As this study tries to use a computational method to generate thumbnails, it is a part of MIR. Additionaliy, as musical salience is considered and the aim is to identify the most representative thumbnails, the study also lies within the field of music cognition. Within MIR, this study falls in the scope of user studies as it tries to understand the need of users to present them the best representation of a song (Futrelle and Downie, 2002). Moreover, hooks could be useful for other categories within MIR: to improve music recommendation, as a way to measure musical similarity, for the generation of satisfying segmentation as they usually start at the beginnings of new sections, and for fingerprinting since hooks can be interpreted as the brains fingerprint of a musical piece (Burgoyne et al., 2013).

Thus, this study focuses on whether the hook of a song could be used to identify the most representative fragment that represents the complete song well enough to make the listeners listen to as little as possible to be able to make a decision. This should hopefully improve audio thumbnails as hooks are inspired by how humans

perceive music and at the same time give extra insight into what hooks and catchiness are.

First, some background into catchiness and audio thumbnailing in general is given in Chapter 2. Thereafter, the method is given in Chapter 3 describing the data preparation, the user study, feature extraction, dimensionality reduction, and the approximation of the representativeness scores for the audio thumbnailing. Finally, the results and discussion followed by the conclusion are given in Chapters 4, 5, and 6 respectively.

Chapter 2

Literature

Before catchiness can be used for automatic music thumbnailing, the notion of hooks and catchiness first need to be defined. The project which inspired the use of hooks for thumbnailing will be elaborated on, along with a related phenomenon, a user study which follows the same approach as this study, and a toolbox for feature extraction to describe catchiness. Thereafter, some insight will be given into thumbnailing and how it differs from relating methods, previous research on automatic audio thumbnailing, and how automatically created thumbnails are evaluated.

2.1 Catchiness

Many listeners can identify easily upon listening whether a song is 'catchy' or not (Burgoyne et al., 2013). They are also mostly capable of pinpointing the catchiest musical segment of a song, the *hook*. However, it remains unclear what catchiness and hooks are, even when these concepts are vital for the understanding of human musical memory. Where Burns (1987) already proposed the analysis of hooks in pop music, Honing (2010) suggests how musical hooks show the potential of cognition-based music retrieval. A common approach within MIR is to use advanced machine learning methods on data to retrieve information from musical data. Honing proposes to let MIR and music cognition elaborate by considering cognitive aspects of the music in the form of hooks. They suggest that using such a cognition-based approach is the next step in improving music retrieval. Therefore, they introduce the notion of a musical hook as the "most salient, memorable, and easy to recall moment of a musical phrase or song" to identify which features affect how music is appreciated, memorised, and recalled.

This definition of hooks and the need for a larger user study to research musical

hooks, and the more general catchiness in music, resulted in the game *Hooked* created by Burgoyne et al. (2013). For this study, they adopted the cognitive point of view of Honing (2010) for the definition of a hook: the most salient, easiest-to-recall fragment of a musical piece. The musical hook here is thus the catchiest part of the song where catchiness is described as long-term musical salience, the memorability of musical fragments. It should be noted that every song will have a catchiest part, even when some music may be catchier or have a catchier hook.

With this notion of musical hooks, Hooked was set up as an experiment to help quantify the effect of catchiness on musical memory (Burgoyne et al., 2013). To be able to reach a much larger number of participants, the experiment was framed as an internet-based game. The assumption for the game is that the easier a fragment is to recall, the catchier that fragment is. The game consisted of three tasks: a recognition task, a verification task, and a prediction task. The recognition task measured how long it took for participants to recognise a tune. In this task the player could listen to a fragment of a song and click on a button as soon as they recognised the song. Here, more points could be obtained the faster the song was recognised. Thereafter, the song would mute and the verification task would follow after which the songs starts playing after being muted for several seconds. While the song was muted, participants were asked to sing along in their heads and afterwards verify whether the song resumed at the correct point. If they identified this correctly, it was assumed that they did recall the song. Lastly, the prediction task was an isolated task to test whether the intuition of the player of what fragments are catchier matched with the formal definition. In this task, participants were asked to choose the more catchier fragment of two of the same songs for extra points.

The results of the game showed that response times in the recognition task could differ up to four seconds (Burgoyne et al., 2013). Further analysis on the obtained data from the experiment was carried out by Van Balen et al. (2015a). The methods and features for this audio corpus analysis are further elaborated on in Section 2.1.3. With their analysis they wanted to identify which attributes predict differences in recognition ratings and what the proposed features model and how they behave. At the same time the study tested how much audio-based corpus analysis tools could add to insight into the data.

For the analysis, Van Balen et al. (2015a) only considered participants that attempted to recognise at least fifteen segments. After the features described in Section 2.1.3 were obtained, a Principal Component Analysis (PCA) was used to reduce the feature-space to a workable number of decorrelated variables. The extracted components were fitted using a linear mixed-effects regression model. This type can handle repeated-measures data which is necessary due to the availability in the data

set of different segments of the same song as well as multiple plays from each participant. Moreover, a linear model is easy to interpret to see how features contribute to the prediction. The results of the study were that audio features are indeed meaningful descriptors. Segments that are easier to recall have a more typical sound and representative segments are more recognisable. As for melodies, those that are recognisable also contain more typical motives and are thus those that are more conventional (Jakubowski et al., 2017; Van Balen et al., 2015a). A role of repetition has also been found in the importance of timbral recurrence. Lastly, a prominent vocal line helped improve the recall in the study. Thus, audio-based corpus analysis tools contribute substantially to insight into how catchiness was measured in the Hooked study.

In conclusion, hooks are framed as the most salient part of a musical piece (Burgoyne et al., 2013), or the fragment of a melody that most people remember or start singing when asked to do so (Honing, 2010). Moreover, representative segments seem to be more recognisable (Van Balen et al., 2015a). Thus, the usage of the musical hook from a cognitive viewpoint could lead to fragments that represent a musical piece to be a good thumbnail.

2.1.1 Earworms

As mentioned, the boundaries between hooks, catchiness, and anything related remains fuzzy (Burgoyne et al., 2013). Two related concepts are hit-song science which tries the predict the popularity of songs and INvoluntary Musical Imargery (INMI). The latter is more commonly known as "earworms" and describes the experience where a tune gets involuntarily stuck in the mind (Burgoyne et al., 2013; Jakubowski et al., 2017; Williamson et al., 2012). While Burgoyne et al. already mentions that INMI are a too narrow definition to be usable for MIR, looking into them might give some insight into catchiness overall. This is strengthened as Jakubowski et al. proposes that the section of a tune that is most easily recalled, the hook, might also be the section that comes to mind involuntarily during an INMI episode.

INMI is an example of internally-directed thought that is not under conscious control (Williamson et al., 2012). It is involuntary, spontaneous cognition and a common everyday experience (Jakubowski et al., 2017; Williamson et al., 2012). Williamson et al. studied the circumstances surrounding INMI episodes; why they happen at any point in time. While the brain activity while experiencing musical imagery is similar to when actual music is heard, musical imagery is experienced without direct sensory instigation and is vivid and veridical. It often consists of repeated fragments and has a link with familiarity in long-term memory and recency.

Positive links between the frequency of INMI episodes and music education and portable music devices have also been found.

Williamson et al. (2012) studied what type of contextual circumstances aid INMI to commence by analysing participant reports. This resulted in eight dominant themes falling within four categories. The first category is *music exposure* which describes recency and repetition. Second is *memory triggers* where the onset of an INMI is not related to recent exposure, but caused by association, recollection, and anticipation. Third are the themes in *affective states* that are associated with the start of INMI episode such as mood, stress, and surprise. The last category describe *low attention states* where the attentional demand is low such as dreams and mind wandering. Of these categories, most INMI reports were linked to recency and repetition in the music exposure category. This highlights the importance of familiarity for the onset of an INMI episode. Concluding, the role of musical memory for INMI is beyond doubt, and recency, familiarity and the omnipresence of music play keys roles in the onset of INMI episodes.

Another study tried to identify what makes a song get stuck in the mind over others (Jakubowski et al., 2017). Both intramusical (musical features and lyrics) as well as extramusical (radio play and context) were examined along with the popularity of INMI. To pinpoint melodic features enhancing memorability the notion of *first-order* and *second-order* features are used which are further described in Section 2.1.2. Thus, Jakubowski et al. tried to predict whether a song would become an INMI based on popularity, recency, and features derived from the melody.

Via an online questionnaire and a data classification method known as random forest, songs that are more likely to become INMI are identified and inspected (Jakubowski et al., 2017). It was found that tunes with a generally faster tempo and a common global melodic contour in comparison to the reference corpus were more likely to become INMI. If the melodic contour does not conform to the norms, song were still likely to become INMI if the melody has a highly unusual pattern of intervals rising and falling. This aligns with the results by Van Balen et al. (2015a) when analysing which features could help predict the scores in Hooked. A possible explanation for the importance of common global pitch contour would be that these would ameliorate the ability to the sing along with the melody easily. However, these findings contradict those of a previous study which indicated the INMI had longer average durations and smaller pitch intervals (see Müllensiefen and Halpern, 2014). Apart from intramusical features, both features related the popularity and recency had significant roles for songs to become INMI. Thus, melodic features, popularity, and recency contribute to the onset of INMI.

Research into INMI shows that some overlap can be found between catchiness

and INMI. Jakubowski et al. (2017) identified the same features that could predict a song's INMI potential as those found when examining the results from the Hooked study (Burgoyne et al., 2013; Van Balen et al., 2015a).

2.1.2 User Study, First-Order and Second-Order Features

The Hooked experiment itself and the analysis of its data was already a user study into what makes certain fragments easier to recall (Burgoyne et al., 2013; Van Balen et al., 2015a). Another example of a study where data obtained from a user study gave insight into the understanding of how music cognition works, is the study by Müllensiefen and Halpern (2014). Müllensiefen and Halpern studied how structural features and context could predict successfully whether a novel melody would be identified correctly by both explicit and implicit memory.

We have an excellent memory for music, but whether memory for music is special is still unknown (Müllensiefen and Halpern, 2014). Müllensiefen and Halpern researched to what extent the features of songs themselves and a frame of reference can predict the memorability of real, but unfamiliar, pop tunes. A discovery-driven approach is taken by obtaining features from stimuli and using statistical learning to decipher which features can predict the behaviour of listeners. This means that first a user study was carried out to obtain data on when listeners explicitly or implicitly recognised a previously heard or novel tune. Thereafter, two questions were to be answered: whether the same features can predict explicit and implicit memory and whether in explicit memory some features or context can drive recognition responses irrespective of a tune being actually heard before. These questions were answered by predicting the memory performance of the participants to identify musical features that could explain the scores found in the user study.

To study this, the assumption that statistical learning also operates in music is needed and that contextual information is used in memory for music (Müllensiefen and Halpern, 2014). Müllensiefen and Halpern propose that even non-musicians can abstract statistical properties from a song and a context. Thus, music is viewed in a frame of reference and compared to context. This gives rise to the earlier mentioned *first-order* and *second-order* features, which have also been used in the studies by Van Balen et al. (2015a) and Jakubowski et al. (2017). First-order features are computed using the intrinsic content of the music or audio itself, such as the average note duration within the melody (Jakubowski et al., 2017; Müllensiefen and Halpern, 2014; Van Balen et al., 2015a). Second-order features reflect the characteristics of the music in context of a corpus. These features thus illustrate how common or highly distinctive a melody is in comparison to a reference corpus. Changing the reference

corpus gives the opportunity to model different cognitive listening contexts.

Apart from first-order and second-order features, Müllensiefen and Halpern (2014) also introduce so-called *summary* features and *m-type* features. Summary features are reductive and summarise the item to one or few values, such as duration and tonality. M-type features are representations of short melodic motives which are tabulated in a frequency table, making them similar to a linguistic n-gram model.

For the study, Müllensiefen and Halpern (2014) used a corpus containing a large collection of pop music and a corpus containing only stimuli presented in the study as reference corpora for the second-order features. Thereafter, Müllensiefen and Halpern tried to explain variability in the results of the user study on the basis of structural features. First, clusters were identified and features were used to predict the results of the implicit and explicit memory task. At the same time, individual features were identified that could have elicited the subjective feelings.

The results indicated that explicit and implicit music memory are moderately correlated (Müllensiefen and Halpern, 2014). The implicit model does use first-order surface characteristics and uses contour and rhythm information. Explicit memory was more dependent on individual encoding strategies and driven by uniqueness, indicating its need for a reference corpus. Summarising, neither draw heavily from first-order features but do rely on statistical features. There was also no difference found between using the full pop corpus as reference corpus compared to using only stimuli presented in the study. Moreover, unusual features made participants think they have heard a tune before even when they have not. On the other hand, flat contours appear novel more often.

Thus, the study by Müllensiefen and Halpern (2014) illustrates an approach using a user study to obtain data from listeners and uses methods to afterwards create models to gain insight into music memory. This is similar to how data from a user study will be used in this study to predict the results using features derived from the audio in hindsight.

2.1.3 Catchy Toolbox

This study worked with the toolbox used by Van Balen et al. (2015a) to analyse the data from Hooked. This CATCHY toolbox[1] was made as Van Balen et al. noted a scarcity of corpus analysis tools for audio data. Corpus analysis describes the analysis of a music collection to gain insight into the music itself. While corpus analysis studies and tools for audio data are limited, symbolic data studies and tools

[1] `https://github.com/jvbalen/catchy`

are more common, such as the work by Huron (2006), the FANTASTIC[2] toolbox (Müllensiefen, 2009), and IDyOM[3] by Pearce (2005) which was based on the work by Conklin and Witten (1995).

With the CATCHY toolbox, Van Balen et al. (2015a) proposes three novel concepts for the study of corpus analysis: three new representations for melodic and harmonic intervals, the use of *second-order* features for audio data, and a definition of song-based and corpus-based second-order features. These concepts are described below to understand what the toolbox has to offer.

First, Van Balen et al. (2015a) describe three new interval descriptors with the purpose of translating simple harmonic and melodic structures to robust representations. The first descriptor is the *Harmonic Interval Co-occurence* (HIC) which measures the distribution of triads in an audio segment. Triads are sets of three notes that can be described in thirds and in this new descriptor, the triads are portrayed by their interval representation. For example, a song with a lot of minor chords will have a strong HIC $_{3;4}$ as a minor triad consist of a minor third with a major third stacked on top. The second descriptor is the *Melodic Interval Bigram* (MIB) which is a three-dimensional matrix that indicates how often triplets of melodic pitches occur less than a certain amount of seconds apart. The last descriptor is the *Harmonisation Interval* (HI) which measures which harmonic pitches in the chroma are accompanied by melodic pitches in the melody.

The second addition by Van Balen et al. (2015a) is the notion of *second-order* features for audio data as described in Section 2.1.2. The inspiration for the use of these features comes from the FANTASTIC toolbox which has these features for symbolic data. The goal of second-order features is to give a context to the values of features, such as whether the value is common or whether the value is rather unique. Second-order features are thus descriptors that reflect on how an observed feature value relates to the corpus it is compared to.

The computation of second-order features differ for audio features in one dimension and those in d dimensions (Van Balen et al., 2015a). Second-order features in one dimension have their typicality described by the log odds that a less extreme value can be observed in the reference corpus. Van Balen et al. proposes to define "less extreme" as "more probable", showing that a density estimation can be made to rank the density for each value of the feature in accordance with the reference corpus.

When dealing with second-order audio features in d dimensions, such as MIB and HIC, two other measures of typicalness are used (Van Balen et al., 2015a). The

[2] http://www.doc.gold.ac.uk/isms/m4s/

[3] https://code.soundsoftware.ac.uk/projects/idyom-project

first measure is Kendall's rank-based correlation which is directly taken from the FANTASTIC toolbox (Müllensiefen, 2009; Van Balen et al., 2015a). The second measure is Information (I), an information-theoretic measure of unexpectedness. Information assumes that the feature itself can be seen as a frequency distribution of observations in the audio and has previously been used as a measure of surprise in IDyOM (Pearce, 2005).

Lastly, Van Balen et al. (2015a) define song-based and corpus-based second-order features. These are based on the notion of expectations that arise from statistical inference by the listener. When a listener hears music, expectations due to familiarity with musical work arise which help to deal with novel music. To approximate these expectations, Van Balen et al. propose to use typicality and surprise in relation to a reference corpus. If this reference corpus consists of a large amount of musical works, expectations with respect to the novel song can be approximated. Van Balen et al. call corpus-based second-order features measures of *conventionality* as they model expectations. Song-based second-order features use only the song itself as reference corpus and thus should indicate how representative a segment is for a song and could to some extent display how much a segment is repeated. The latter type of second-order features is dubbed *recurrence* by Van Balen et al.

2.2 Music Thumbnailing

As discussed, a music or audio thumbnail is a continuous audio segment that best represents a musical piece (Chai and Vercoe, 2003; Cooper and Foote, 2002; Huang et al., 2017; Müller et al., 2013). However, the concept of a music thumbnail is generally ambiguous (Müller et al., 2013), since it could be described as a characteristic segment (Levy and Sandler, 2006), a key part (Schuller et al., 2008), the main tune (Nawata et al., 2011), a down-sampled version (Bartsch and Wakefield, 2005), or the chorus of a pop song (Huang et al., 2017; Schuller et al., 2008).

As music thumbnails are shorter representations of musical pieces, they can be used as an effective way to navigate through large music collections with tracks of possible interest (Cooper and Foote, 2002; Levy and Sandler, 2006). Thumbnails thus provide the listener a quick impression of a song (Levy and Sandler, 2006) and could be used for web browsing, web searching, and music recommendation (Chai and Vercoe, 2003; Levy and Sandler, 2006; Schuller et al., 2008).

It is also important to distinguish thumbnailing from other methods such as summarisation and fingerprinting. Earlier research into creating audio thumbnails use the terms thumbnailing and summarisation interchangeably (Chai and Vercoe, 2003; Cooper and Foote, 2002). While these studies still aimed to detect a continuous

segment within a song, the two terms should not be confused. Summarisation is the creation of a shorter fragment to represent a musical piece by combining snippets of all different parts (Cooper and Foote, 2002; Silva et al., 2018). This means that a summary is made with parts dispersed over the whole song which are then combined to create a reduced representation. This differs from audio thumbnailing as the aim is to find a continuous segment of the song without dissecting song sections.

Another term that is sometimes confused with audio thumbnailing is audio fingerprinting. In a review on audio fingerprinting, Cano et al. (2005) describes an audio fingerprint as a compact content-based signature that summarises the audio recording. Important to notice is that a fingerprint is a much shorter numeric sequence, instead of an audio fragment. Fingerprinting extracts the perceptual digest of a musical piece by using acoustic relevant characteristics of the audio (Cano et al., 2005; Van Balen et al., 2015b). Commonly, fingerprints are stored in a database and used to identify new data, which has its characteristics calculated and matched against the data base. This means the fingerprinting allows for the identification of audio independently of format and meta-data. It differs greatly from audio thumbnailing as the summary is a sequence or vector and not used as representation interpretable for human listeners.

This research focuses only on finding a good music thumbnail: a continuous segment that can represent a musical piece. The following subsections discuss several previous studies on audio thumbnailing.

2.2.1 Previous Research

The definition of what a good thumbnail is differs and has an influence on the approach taken in studies to create thumbnails. A common assumption in these studies is that the best thumbnail is the segment that is repeated most often (Chai and Vercoe, 2003; Müller et al., 2013). This belief causes research to focus on finding the structure of an audio file to detect the most repeated segment.

The first example of an automatic thumbnailing method that tries to detect the most repeated segment is by Müller et al. (2013). They argue that a typical thumbnail would be a segment that has many repetitions that cover large parts of a musical recording. An example would be the chorus of a pop song which is repeated often and covers a large amount of the song. However, problems may arise due to variations, which could make segments that are considered repetitions of earlier segments differ significantly. To identify a segment with many (approximate) repetitions, Müller et al. introduce a fitness value that expresses how well a segment can "explain" the repetitive structure and use self-similarity matrices to check what amount of the

music recording is covered by related segments.

Chai and Vercoe (2003) also mention that it is still unclear what makes a specific fragment the most memorable or distinguishable and assume for this study that this is true for the most repeated part of a song. Apart from using the most repeated segment, Chai and Vercoe also suggest the option of choosing a fragment containing a transition between different sections of the song to give an overview of the complete work, and a strategy similar to summarisation where phrases from different sections are stitched together. To create an audio thumbnail, Chai and Vercoe start by identifying the recurrent structure and then use one of these three strategies to select a thumbnail.

Bartsch and Wakefield (2005) start by doing beat-synchronous segmentation to obtain frames per beat. Thereafter, chroma-based features are computed per frame and these features are used for the calculation of correlation between frames. The audio thumbnail is chosen by checking for the most similar pair of audio fragments and selecting the earlier fragment of this pair. Bartsch and Wakefield assume that the most similar pair of fragments will consist of the chorus as that is the most likely part to be repeated as similar as possible. However, they do note that the features used are derived from the musical signal and is not motivated by the perception of the listener.

Similarly, Cooper and Foote (2002) assumes the best thumbnail to be the segment that is the most similar to the average sound. Their approach was to find the segment with the highest similarity to the entire musical piece. This approach was thus very dependent on finding the most repeated part. Cooper and Foote also suggests that the best thumbnails will begin and end at meaningful segment boundaries.

Levy and Sandler (2006) also notes that music has structural sections and how these sections could be labelled per section type to identify the most repeated section type. First, they start with segmentation to divide the song into its sections. Each section is thereafter labelled and the most common section type is thereafter identified. Lastly, the second segment of the most occurring section type is chosen as a thumbnail as Levy and Sandler assume that the middle of a song is more representative.

Apart from searching for the most repeated part of a song to use as a thumbnail, another common approach is to detect the chorus. This is still similar as can be seen in the example given by Müller et al. (2013) which mentions that the chorus of a pop song chorus is often the most repeated part. Similarly, Levy and Sandler (2006) also noted that the chorus of pop songs are often used for thumbnails. Schuller et al. (2008), for example, belief that repeated sequences such as chorus sections are the most mnemonic parts and therefore tried to extract the chorus from a song as

thumbnail. Akin to Müller et al. (2013), Schuller et al. create similarity matrices where diagonals in the matrices correspond to similar segments. To automate the detection of the diagonals, a computer vision technique to detect edges in images is used to extract bright diagonals and determine segments of interest. The selected audio thumbnail is the best remaining segment in regard of its mean similarity. Their results indicate that there is a difference per genre and that thumbnails that were evaluated incorrectly also contained characteristic parts of the songs.

A more recent study by Huang et al. (2017) also proposes that chorus detection may lead to good audio thumbnails. They discuss how the chorus can be considered the most memorable part of a song. To detect the chorus of a song, Huang et al. propose to use emotion recognition software as they assume that the main function of music is to communicate emotion. For the emotion recognition, a neural network model is used with which they test whether a music emotion recognition model reveals anything about the structure of the song. The model first detects the emotion of segments of the song. The emotion cluster with the largest amount of segments is considered the chorus and from these segments, the one closest to the middle of the song is chosen as thumbnail. The results indicate that for the thumbnails of 80 songs, 50% overlaps with the chorus sections.

Lastly, Nawata et al. (2011) use a different approach for automatic thumbnail generation. They call a music thumbnail the main part of the song and use the activation of audio objects and their location to detect sections. By analysing the activation, they try to detect structural changes and try to identify the main composition section to select as a thumbnail.

Concluding, research into automatic music thumbnails is scattered in approach but mostly assumes the best audio thumbnail is either the chorus of a pop tune or the part that is repeated the most. This research adds a new insight into the field by trying to capture the perception of the listener.

2.2.2 Evaluation

Another challenge with automatic music thumbnailing is evaluating the created thumbnails. Previous research has mostly compared the generated thumbnails to manually annotated thumbnails (Bartsch and Wakefield, 2005; Müller et al., 2013; Nawata et al., 2011; Schuller et al., 2008). Müller et al. evaluate by checking overlap between the boundaries of automatic generated thumbnails and a manually generated ground-truth annotation, and computing the F-measure afterwards. The F-measure, precision, and recall are also used by Nawata et al. who use it for their "objective" evaluation by checking the timing-detection accuracy in accordance with a manu-

ally given answer. Likewise, Bartsch and Wakefield also use precision and recall as evaluation scores to test the overlap between generated thumbnails and thumbnails selected by a single listener. Schuller et al. compared manual annotation to their generated ones by checking whether the initial positions of the thumbnails matched within in a certain threshold.

The use of manual annotation to compare the results of a new audio thumbnailing algorithm seems to be the most common. However, Chai and Vercoe (2003) propose several criteria to evaluate music thumbnails with based on properties and previous experiments with human listeners. The evaluation is done by considering four criteria: the percentage of the thumbnails containing vocals, the percentage of the thumbnails containing the song's title as they assume this aligns with the hook of the song, the percentage starting at the beginning of a structural section, and the percentage starting at the beginning of a phrase.

However, the problem with comparing to manual annotations is that music annotation is subjective (Koops et al., 2017). Koops et al. noticed the ambiguity in chord annotations in, for example, online repositories containing multiple versions of annotations for the same popular song. Therefore, one single annotation may give rise to problems as different listeners may disagree. This could also be the case for audio thumbnailing as different annotators might think other segments are better representations of the same song.

Koops et al. (2017) themselves propose a method to tackle the ambiguity between different annotations for chords by modelling subjectivity and personalising chord labels for each annotator. They showed that multiple reference annotations outperformed single reference annotations. They also discuss a study by Burgoyne et al. (2011) which uses several opinions to create an optimal "mean" annotation. Nawata et al. (2011) also uses several listeners for a "subjective" evaluation by letting them score whether a possible structural change can be heard in a presented segment. This aligns with the evaluation used in this study which uses mean user annotation obtained via a user study to evaluate computed audio thumbnails.

Most previous studies also do not compare to baselines and setting such baselines may pose new problems. Müller et al. (2013) propose the use of two baselines to compare their generated thumbnails with: the entire song and a thumbnail starting at the second sixth part of the song.

For this study, the idea of subjectivity in individual reference annotations is a reason to use a "mean" annotation which is obtained via a user study. The user study will thus lead to subjective score for different segments on their representativeness.

2.3 Hypothesis

The aim of this research is to create a new cognition-inspired music thumbnailing method. This is achieved by doing a user study where listeners rate fragments on their ability to represent the full song. The set-up of the user study and the features used are inspired by research into catchiness and INMI. This aligns with thumbnailing as the cognitive definition of catchiness describes the long-term salience of musical fragments (Burgoyne et al., 2013; Honing, 2010), which could also measure representativeness (Van Balen et al., 2015a).

Thus, the first questions that arise are: can the notion of catchiness be used to create cognitively inspired music thumbnails? And can the representativeness of a fragment be described with features related to catchiness and hooks?

The expectation is that with enough participants in the user study, the scores of fragments differ enough to be able to identify meaningful features to explain them. If long-term salience also measures representativeness, the same features found in previous studies related to catchiness and INMI should be identified. This would mean that a general higher commonality and recurrence of a fragment with regards to the reference corpora should align with fragments obtaining higher scores in the user study (Van Balen et al., 2015a). This means that a more typical sound, repetition, a prominent vocal line, and conventional melodies (Jakubowski et al., 2017; Van Balen et al., 2015a) benefit thumbnails. Thus, it is expected that these features which are measurable with the catchy toolbox can approximate the user study scores. This would lead to a new music thumbnailing method.

Another theory is that hooks start at the start of structural sections (Burgoyne et al., 2013). Similarly, Cooper and Foote (2002) suggest that the best thumbnails start and end at a meaningful boundary. While the latter is not relevant in this study as Muziekweb uses a fixed length for previews, testing the influence of starting at a structural boundary might benefit thumbnails. Therefore, the hypothesis is that using boundary detection for segmentation benefits the generated thumbnails.

Chapter 3

Methods

This study aims to approximate the representative power of fragments to select the most representative fragment as audio thumbnail. The representativeness of fragments is rated via a user study in Dutch similar to the prediction task in the Hooked study (Burgoyne et al., 2013). Afterwards, features per segment are computed with the CATCHY toolbox (Van Balen et al., 2015a) which are used to explain the user study scores of the segments. Lastly, a "catchy" function is created as a model of how human listeners score the segments to derive the most representative thumbnail per song.

3.1 Data

The methods in this research are fine-tuned on popular music as previous research on catchiness has focused on the same type of music (Burgoyne et al., 2013; Müllensiefen and Halpern, 2014; Van Balen et al., 2015a). Schuller et al. (2008) already noted that the preferred thumbnails differ per genre, which implies that the results will not be suited for any other genre than pop music. Furthermore, as the aim of the new thumbnailing approach is also to be used by Muziekweb afterwards, the used audio is provided by them and consists of a selection of 60 pop songs which are listed in Appendix A. For each song, six ten-second segments are selected to be evaluated in the user study.

3.1.1 Selection of Music

The selection of music for this study is based on statistics derived by Muziekweb. The data that was used were two lists of the 100 most listened to songs on Muziekweb

in 2017 and 2018. This list is thus fine-tuned on the data coming from the behaviour of Muziekweb users. First the counts per song for these two years were summed up, giving a list of 121 unique songs. An interesting feature of this list is that a large amount consists of French and German songs. The reason could be that Muziekweb is the first hit via Google search when either French or German songs are queried. However, as these songs are less likely to be known by the general Dutch population, it was opted for this study to only use one song per language in a different language from either Dutch or English. This choice should help with recognition during the user study. However, one extra song in German was included as the title of that song is in English and was overlooked.

Thereafter, another decision was made for the remaining songs, removing double artists and albums to keep the remaining music as diverse as possible. If an artist or album had several listings, the song with the highest listening count was included. An exception was made for songs appearing on the same compilation albums of "greatest hits". While compilation albums may skew the data, the songs may still be considered representative of general pop music of different time periods while showing the diversity at the same time. A last song that was excluded as an exception was a live recording of how the band was introduced.

The final selection of used songs in this study can be seen in Appendix A. All the songs in this list were provided by Muziekweb in FLAC.

3.2 Frame Segmentation

The rating in the user study is done on a selection of ten-second segments, or excerpts, from the original audio. Hooks mostly occur at the start of structural sections (Burgoyne et al., 2013; Honing, 2010), meaning that segments starting at these sections make more sense to assess. In Hooked, Burgoyne et al. used the Echo Nest as a structural segmentation algorithm, which is not available anymore to use. Therefore, here the identification of these beginnings for segmentation is done via an algorithm available in the Python package called MSAF (Nieto and Bello, 2015).[1]

MSAF, or Music Structure Analysis Framework, is a framework which facilitates the analysis, evaluation, and comparison of music structure analysis algorithms (Nieto and Bello, 2015). Music Structure Analysis (MSA) aims to detect boundaries, the exact points in a musical piece where sections start or end, and aims to do structural grouping by classifying the identified segments into groups. The idea behind MSAF is to create a transparent framework to compare different algorithms for MSA and

[1] https://msaf.readthedocs.io/

already includes several implementations of MSA algorithms.

For this study, one specific implementation from MSAF is used for boundary detection. This is the algorithm based on the original publication by Serra et al. (2014). This algorithm aims to annotate songs by looking at structure features and time series similarity. The structure features encapsulate global characteristics which in combination with local measurements results in the estimations of segment boundaries. This algorithm is chosen as it is the standard algorithm used in the framework while still yielding feasible results in comparison to other algorithms in MSAF . However, MSAF does not use the original implementation by Serra et al. (2014) as this was not publicly available (Nieto and Bello, 2015). At the same time, the implementation in MSAF is not discussed by Nieto and Bello (2015), which makes it impossible to compare this boundary detection algorithm to others.

A second choice that has been made is which type of time series is used for boundary detection. MSAF has the option to use several types based on Librosa[2] implementations, namely tempograms, tonal centroid features (Tonnetz), Pitch Class Profiles (PCP), Mel-Frequency Cepstral Coefficient (MFCC), and a Constant-Q Transform (CQT) scaled to a dB-spectrum. Considering that the features in the catchy toolbox are mostly harmonic, tempograms for rhythmic content and MFCCs for timbral representations do not seem interesting for boundary detection. From the remaining three options, the Tonnetz implementation resulted in far too few boundaries to be feasible, while the CQT implementation was very slow. This leaves the PCPs as last option to use which aligns with previous studies describing the importance of chroma's (Chai and Vercoe, 2003; Schuller et al., 2008). The PCPs as used in MSAF describe harmonic content using Librosa's implementation for CQT chromagrams.

Thus, for each of the 60 selected songs, boundaries are detected based on the CQT chromagrams using the implementation in MSAF of Serra et al. (2014). First, each song was converted to WAV to ensure compatibility with the packages in Python. Using MSAF, a list with times per song is given of detected boundaries. For each of these boundaries, a new WAV-file was created that lasts 9.95 seconds which starts at the identified boundaries. Previous studies have shown that ten seconds is plenty of time for the listener for recognition (Burgoyne et al., 2013). Most listeners can recall a catchy song from memory after only 400 milliseconds. Also, using shorter fragments should lower the computation of features per fragment later on in the study. Moreover, Muziekweb mentioned that due to copyright they are only allowed to display a maximum of 29.9 seconds per song. As in the user study three different segments per song will be displayed, choosing to use segments with a length of

[2] `https://librosa.github.io/librosa/`

9.95 seconds prevents possible copyright issues. Boundaries were ignored when the resulting segment exceeds the end of the song. Per song this resulted in multiple excerpts with a total of 737 for the full corpus in separate WAV-files. To get an impression of the detected boundaries, the resulting excerpts were examined. The fragments do not start perfectly at the different sections of the song, but do seem to have captured the sections well. This would mean that the start of each excerpt is not as well as hoped for, but the general sense of the different sections of the songs is captured with this method. Moreover, as thumbnails generally use fade-ins and fade-outs, the exact starting point of the fragment becomes less important.

3.2.1 User Study Fragments

For the user study, a selection of six excerpts per song is made to be evaluated. This is done for the 60 songs to be studied as well as two additional songs which are used as examples in the user study. Four of the excerpts per song were those that start a structural boundary as detected by the algorithm in MSAF. To ensure there is no bias within the chosen excerpts, the selection of excerpts is done completely at random. This might cause some of the excerpts starting at structural boundaries to be less representative as the intro of a song may be chosen for example. Previous audio thumbnailing studies assumed the middle of a song to be the most characteristic (Huang et al., 2017; Levy and Sandler, 2006), which might not be a part of the chosen excerpts due to randomisation. There are, however, songs where the intro does serve as the hook (Burns, 1987; Kronengold, 2005), which fortifies that random selection of fragments is preferred. After listening to the selection of excerpts starting at structural boundaries, there does seem to be a preference for in the selection for fragments starting at the intro, which might influence the results.

The remaining two excerpts for the user study per song are used as base cases. The first base case is how previews are currently generated by Muziekweb, which is by choosing a segment completely randomly. The second base case is a rather arbitrary one-minute-in segment. The latter base case should give a more reasonable excerpt as the song should have commenced into a more well-known part, skipping the intro, but keeping distance from the end where variations of the different sections may occur. However, the start of the segment could be rather abrupt as it starts at such an arbitrary point. But also for the base cases, the standard use of fade-ins and fade-outs in thumbnails reduces the problems of fragments starting abruptly.

3.3 User Study

The aim of the user study is to provide scores of the representative power for each of the selected six segments of the 60 songs. This is inspired by Müllensiefen and Halpern (2014) who used a user study to afterwards identify features and context that could explain the trigger of recognition of novel melodies. The scores in the current study are obtained via an online survey where participants can rank the ten-second segments in threes. Afterwards, the user study results are used to obtain scores of representativeness which are approximated with a catchy function to obtain the best segment as music thumbnail.

3.3.1 Participants

The user study is a web-based experiment where participants are reached via social media. A link to the experiment was placed on the homepage of Muziekweb, and the Twitter and Facebook-page of Muziekweb. It was accessible from April the 3rd, 2019 up to May the 27th, 2019. There is no expectation that Muziekweb-users differ from the general population and therefore the link has also been distributed by the researchers.

The participants should have been at least 18 years old and a check was made before the actual experiment where participants could confirm their age and give their informed consent (see Appendix B). The expectation was that only healthy participants of the general population participated.

The estimated amount of participants that was needed to find a significant result is approximately 450. This amount was estimated as follows. In this study, six ten-second segments per song are to be rated. All these segments are assessed in pairs, which leads to fifteen combination of pairs per song. The general rule-of-thumb is to have 30 measurements per pair, which leads to 450 assessments. As participants are asked to score three segments simultaneously, each song needs to be evaluated 150 times. However, participants are expected to score 20 songs approximately instead of the full data set of 60 songs used in this study due to the duration of the full experiment. This leads to a total of 450 participants.

The expectation of participants scoring 20 songs leads to an expected duration of the experiment of 15 to 20 minutes. The scoring of one song should take approximately 45 seconds to 1 minute. However, participants may continue to score more songs or stop at any given moment.

3.3.2 Survey

The user study as presented to the participants was an online survey. The main task was similar to the prediction task in the study by Burgoyne et al. (2013) where participants were asked to choose which of the two presented fragments of the same song was perceived as being catchier. Similarly, in this study, participants ranked the representativeness of three presented fragments of the same song.

The study here consisted of a page for each song which displays three fragments as playable audio-files along with the title and the artist of the song to help their memory. The participants were asked to rank the three fragments on how well they helped them to get a sense of what the song is ("*een idee van het nummer*"). This phrasing should have triggered a gut feeling of their impression of the song. The ranking was done by ordering the fragments on their capability of conveying the sense of the song. At the same page, the participants were also asked to indicate whether they are familiar with the song or not. If they were, it was assumed that they had a clearer notion of what segments are more representative or memorable. If they did not know the song, the ranking of the segments could give an indication of whether the listener gets the feeling of understanding what the song is about. The hypothesis was that to convey a sense of what the song is about does not rely on familiarity which is evaluated via this question.

The survey started with a page that gave some background on the study, followed by the informed consent. Participants were notified that they can stop the experiment at any moment as participation is completely voluntary, that the obtained data is used for research and stored securely, and that they give their consent to participate (see the informed consent text in Appendix B). By checking a box on the page they confirmed that they participate voluntarily and that they were at least 18 years old. Only after this confirmation did the actual study start.

3.3.3 Implementation of the Survey

The implementation of the user study is done via Qualtrics,[3] an online survey platform to facilitate building surveys and keeping track of the results. This platform was preferred to ensure that displaying music fragments is possible, to ease the procedure of putting together the survey, and to ensure safety of the music and obtained data. At the same time, Qualtrics had a great influence on the final design of the survey which is discussed below.

The tool in Qualtrics to create surveys has a distinction between blocks and

[3] `https://www.qualtrics.com`

questions can be made: each block can consist of multiple questions and is displayed as one page with a customisable next-button. For this study, this means that each song is put into one block with two questions: a question to rank three segments of a song and another question to check whether the participants are familiar with the song.

In Qualtrics, there are four options to display rank-order questions: "drag-and-drop", radio buttons, text boxes, and select boxes. The select boxes are the only rank-order type questions that does not support music players. From the three remaining options, the drag-and-drop question type is chosen to rank the fragments as this type is the most mobile user friendly; text boxes expect the respondent to type their preferred ranking and radio buttons do not fit the average screen without scrolling vertically and selecting a ranking. Moreover, moving items in a preferred order seemed more natural to represent mental ranking.

The drag-and-drop question type asks respondents to drag the options in their preferred order. In this study, the option that was put at the top should be the most representative fragment according to the participant. A limitation within this type is that if one of the options is not picked up at least once, the ordering of the items by the respondent are not recorded. Therefore, it was chosen to force participants to at least drag one item once by giving a built-in warning message if nothing has been moved. While this may annoy participants during the experiment, this ensures that the results are stored in Qualtrics. Additionally, Qualtrics also displays that a ranking is recorded by adding numbers in front the items after the respondent has moved at least one item as shown in Figures 3.1 and 3.2. This might make it easier to know whether the response is recorded during participation.

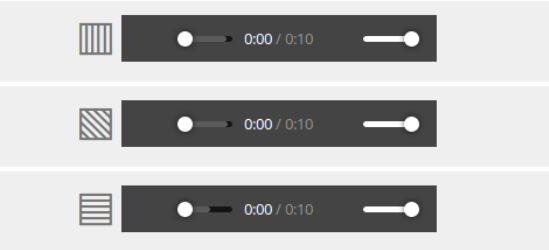

Figure 3.1: Drag-and-drop rankings as displayed in Qualtrics when the participant has not moved any fragment

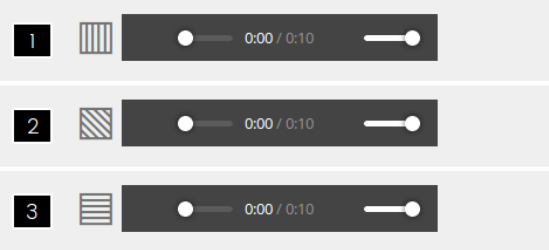

Figure 3.2: Drag-and-drop rankings as displayed in Qualtrics when at least one fragment has been moved at least once

Another problem with the drag-and-drop question type is the insertion of fragments and how these fragments can be distinguished between themselves when they

are moved. To be able to present the fragments in players via Qualtrics, all fragments need to uploaded to Qualtrics and can thereafter be inserted into the questions one by one. This means that all the fragments are available on the Qualtrics servers and that they can only be played via the nonadjustable player available in Qualtrics. To still be able to distinguish between the nonadjustable players without using text or numbers which could lead to bias, the chosen solution here was to use Unicode symbols to differentiate between fragments. The chosen symbols can be seen in Figure 3.3 and consist of six differently filled squares which do not have an apparent ordering themselves. These symbols were added in front of the players (see Figures 3.1 and 3.2, and Appendix C).

Figure 3.3: Unicode characters used in the survey to differentiate between fragments

To ensure that the bias of the participants per fragment in the overall study is minimised, the songs and fragments per song are randomised via the built-in randomisers in Qualtrics. This means that the musical pieces that are actually tested are ordered completely random, as well as which three fragments are displayed per participant and in what order. As the song order is randomised, this should also outweigh the amount of rankings per fragment per song when participants do not finish the full survey.

Lastly, Qualtrics also had an impact on more general choices made for the survey. First, Qualtrics provides the option to prevent "ballot-box stuffing" by not allowing users to participate from the same IP-address more than once. This was selected to ensure that participants only participate once as ranking the same song twice could easily lead to bias. Second, Qualtrics does provide the option to display a progress bar, but it was chosen to ignore this feature as this might discourage participants due to the duration of completing the survey. By not displaying a progress bar, hopefully, participants only rank the amount of songs they feel comfortable with. This does also add a limitation by discouraging participants that want to know what they are up to when participating. Third, the survey ends with an question whether the participant wants to receive the final report of this study afterwards and the standard Qualtrics thank-you message. Qualtrics did not have an option to add a button to every page which would automatically abort the survey, which means that only participants that complete the full survey get to see this question and message. Therefore, it was chosen to add thank-you messages to the introduction. Still, this means that

participants interested in the results who have not finished the complete survey are not able to give their email-address to receive the final version. Last, Qualtrics has the option for participants to continue their survey after closing the survey tab if wanted. However, while Qualtrics waits for respondents to continue, it does not record that made progress. Thus, here it was chosen to let Qualtrics automatically close the survey after seven days of no activity. This should give participants the possibility to take their time for the survey while still ensuring the results are stored by Qualtrics after a reasonable amount of time.

Appendix C shows screenshots with translations of the several parts in the survey as displayed in Qualtrics.

3.3.4 Storage of Data

Qualtrics stores obtained results according to the ethical laws of the university and thus it is safe to keep the results on Qualtrics. The data obtained via the survey can be downloaded in csv-file format with UTF-8 encoding, which can be used in the following steps of this study.

The results that are obtained via Qualtrics are the rankings of the three fragments per song for each participant. The format of the file is to show the questions and the possible options as header for the columns, while the numbers of the ranking ("1", "2", "3", or nothing when the fragment was not ranked by the participant) are shown in rows per participant. To ensure that the results could be easily traced back to their respective songs and fragments, invisible text in HTML was added to the survey to add Muziekweb's song-ID and the start of the second in milliseconds. This invisible text does not influence the look of the survey apart from slightly larger spacing, but does add the text to the csv-file to identify the songs and their fragments.

One problem with the downloaded results from Qualtrics happens due to the prevention of ballot-box stuffing. This ensures that participants cannot participate more than once and is tracked by Qualtrics via IP-addresses. These addresses are not visible in the csv-file when the participant has finished the survey, but are visible for those who have not finished the complete survey. To ensure anonimity to participants, the row containing this information is immediately manually deleted after downloading the results from Qualtrics.

3.4 Feature Extraction

Both the extracted segments with MSAF and the rankings obtained via Qualtrics are not immediately usable. First, the features that are used to describe the segments are

computed with the CATCHY toolbox by Van Balen et al. (2015a) are discussed and explained. Then, a method to transform the rankings of the user study in a "worth" per segment via a model called Plackett-Luce is given. Lastly, a computation is given to obtain a familiarity score as additional feature. By computing these features, descriptors of the sound, as well as measurable values based on the user study are obtained which are used in the approximation steps of this study.

3.4.1 Catchy Features

The derivation of the catchy features for each segment separately is done with the CATCHY toolbox[4] which was introduced in Section 2.1.3. This toolbox can compute several first and second-order features to describe the audio signal (Van Balen et al., 2015a). Afterwards, these features can be used to approximate the survey results.

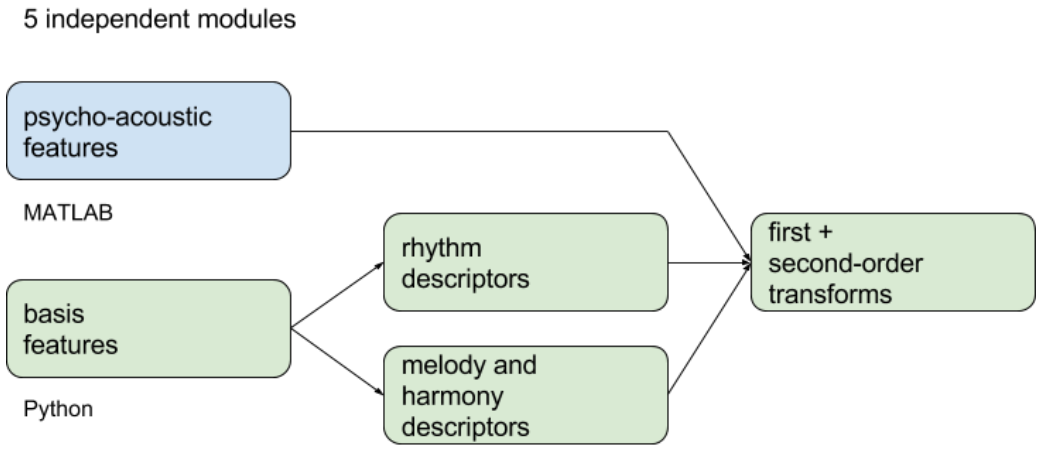

Figure 3.4: Modules in the catchy toolbox by Van Balen et al. (2015a)

To understand how the different first and second-order features are derived from the audio, an overview of the modules in the CATCHY toolbox can be seen in Figure 3.4. Each module computes a different set of all the features. The first module called "psycho-acoustic features" computes features with the MIRtoolbox[5] in Matlab.[6] The features in this module are loudness (mean and standard deviation), roughness (mean), and sharpness (mean). Loudness can be described as the sound intensity of the stimulus as perceived by human listeners (Van Balen, 2016; Weihs

[4] https://github.com/jvbalen/catchy
[5] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox
[6] https://nl.mathworks.com/

et al., 2016). Thus, a higher value on loudness would mean that the listener perceives the sound as being louder or having a greater volume. Roughness in music occurs whenever a pair of sinusoids are close in frequency. A higher roughness can be perceived as pitch perception being unclear, beats less defined, and a generally noisy sound. This can for example occur when two dissonant notes are played causing the perception of a rough sound. Lastly, sharpness of a tone is measured by testing whether the centroid of the sound spectrum is in the higher end. A higher sharpness means that the sound has more power in higher frequencies, causing a shriller sound. After computing these features in Matlab, the results can be read in Python to merge with the other modules.

The other modules in catchy toolbox are all computed in Python. The "basis features" module calculates standard MIR features derived from other toolboxes, namely: Librosa's MFCCs (total variance), estimates of the melodic pitch heights from Melodia[7] (mean and standard deviation), and chroma from HPCP.[8] MFCCs, short for Mel-Frequency Cepstral Coefficient, describe some part of musical timbres in terms of bands, with each band describing a specific aspect (Van Balen, 2016; Weihs et al., 2016). Each band is computed with a different coefficient and often thirteen bands are sufficient to capture audible timbral aspects. However, separately these bands cannot be interpreted apart from the first band agreeing with energy and loudness. Therefore, some kind of mean value cannot be derived. The estimates of the melodic pitch heights are a method to approximate the melody in the audio signal. Lastly, the chroma feature is a mapping of the energy for each of the twelve tones in the Western tonal system over time. This is the only feature that is purely computed to be used for the computation of other features and is not a part of the features used in the following steps in this study.

The features of the "basis features" module feed into the "melody and harmony descriptors" module which uses these features to compute the higher-dimensional features introduced by Van Balen et al. (2015a) (see Section 2.1.3): the Harmonic Interval Co-occurence (HIC), the Melodic Interval Bigram (MIB), and the Harmonisation Interval (HI). As a short recap, the HIC describes the distribution of triads in an audio segment based on their interval representation. The MIB indicates how often triplets of melodic pitches occur in the audio. Lastly, the HI measures which pitches in the chroma are accompanied by pitches in the melody. These higher-dimensional features are a more sophisticated way of depicting harmonics in the audio signals, which means that the computed chroma feature can be ignored.

Figure 3.4 also shows a module called "rhythm descriptors" which is still under

[7] https://www.upf.edu/web/mtg/melodia

[8] https://www.upf.edu/web/mtg/hpcp

development and is supposed to include Inter-Onset Intervals (IOIs) and a beat-tracker. As it is still unavailable, it was chosen for this study to not include rhythmic features.

The last module "first and second-order transforms" uses the computed features of the other modules to calculate the first-order and the corpus-based and song-based second-order features (see Section 2.1.3). As explained in Section 2.1.2, first-order features describe intrinsic content of the audio itself. Normally, this is done in terms of averages and standard deviations as is shown for the features in the catchy toolbox in Table 3.1. Second-order features describe the commonness or uniqueness of the feature values in comparison to a reference corpus. Thus, corpus-based second-order features indicate whether the values are common within a corpus of several songs, while song-based second-order features show the commonness within the song itself.

Module	Feature type	First-order	Second-order
Psycho-acoustic	Loudness	Mean	Ranked odds
		Standard deviation	Ranked odds
	Roughness	Mean	Ranked odds
	Sharpness	Mean	Ranked odds
Basis	MFCC	-	Independent log odds
		Standard deviation	Ranked odds
	Melodic Pitch height	Mean	Ranked odds
		Standard deviation	Ranked odds
Melody and Harmonic	HI	Normalised entropy	Ranked odds
		-	τ
		-	Information
	HIC	Normalised entropy	Ranked odds
		-	τ
		-	Information
	MIB	Normalised entropy	Ranked odds
		-	τ
		-	Information

Table 3.1: Overview of the different derived features with the cathy toolbox (Van Balen et al., 2015a)

Thus, the last module computes the first-order features by calculating the averages and standard deviations of features derived in the previous modules. An exception is made for higher-dimensional features which are described by normalised entropy. The second-order descriptors for the features computed in the psycho-acoustic and basis modules are computed by taking a look at the rank of each feature values probability density estimate in comparison to the corpus, normalising the rank by the number of observations, and taking the log odds (Van Balen et al., 2015a).

This is also done for the the normalised entropy descriptors of the HIC, MIB, and HI. Additional to these features, two different second-order measures are used which are given by computing Kendall's and Information, which is information-theoretic measure of unexpectedness. The only exception are the MFCCs, which are multivariate with independent features, where the mean of the first thirteen components is used to derive an additional second-order feature similar to those of the basis features by considering each component as a one-dimensional feature. Table 3.1 gives an overview of all the first and second-order features that are obtained per feature type.

3.4.2 Exploratory Factor Analysis and Principal Component Analysis

The aim is to use the first-order and second-order descriptors which describe intrinsic characteristics and commonness of values respectively in the following steps in this study. Similarly to Van Balen et al. (2015a), the dimensionality of the features is reduced after their derivation. Van Balen et al. used a Principal Component Analysis (PCA) for dimensionality reduction with the aim of identifying features that explain the same source of variance in the data.

In this study, the main method for dimensionality reduction is Exploratory Factor Analysis (EFA). In contrast to PCA, EFA only regards shared variance and considers latent variables causing an underlying structure (Osborne et al., 2008). In PCA, components are computed using all of the variance of the features for its solution. Thus, EFA becomes a preferred method to ensure no variance is inflated as only shared variance is considered, while the interpretability is improved as an underlying structure via shared variance is examined. However, as Van Balen et al. (2015a) used a PCA, it is also evaluated in Appendix F as comparison to the model obtained via EFA and their study.

The dimensionality reduction is done via the Psych package in R (Revelle, 2011). For the reduction, the feature values of all obtained fragments is used. Thus, the additional fragments obtained via MSAF that were not selected for the user study are also used here to have more occurrences of the features.

The process of reducing the dimensionality has several steps. The first one is to obtain the correlation or covariance matrix of the features. The type of this matrix is dependent on whether the features are normally distributed. Then, selecting the amount of dimensions to map the features onto needs to be chosen. Both too many as not enough factors can have harmful effects on the results (Osborne et al., 2008). This is done via a test available in the Psych package that plots four different methods

to determine the amount of dimensions: Very Simple Structure (VSS), complexity, emperical BIC, and root mean residual. Each of these methods gives a score per amount of factors in a model according to the feature space. Generally, one would chose the factor where the VSS for "1" is the highest, the complexity remains as low as possible, the BIC has reached its lowest point, and the root mean residual has an elbow point. However, one optimal method to determine the amount cannot be found. To gain more insight, the clustering algorithm called Item Clustering Analysis (IClust) is also run to gain insight into its preferred amount of clusters.

After choosing the amount of factors or components, the modelling can be done via EFA. The model is created with "Varimax" rotation, the most common choice for dimensionality reduction where factors are uncorrelated and results are easier to interpret (Osborne et al., 2008). The last choice made in this process is to use minimum residuals as factor extraction method. Minimum residuals is a method used for exploratory and descriptive analyses (Tinsley and Tinsley, 1987). This method assumes both the subjects and features to be populations and therefore, generalisations can not be made.

After creating a model with these choices for the EFA, the model can be used to compute the values per factor for each of the segments. These are the new "features" to describe the fragments, which are used to explain the user study results.

3.4.3 Segment Worth

The rankings obtained in the user study have a format of displaying which of the fragments were ranked as first ("1"), second ("2"), and third ("3"). The unseen fragments do not obtain a score. This means that the results are a ranking of a discrete or qualitative choice where there are no continuous variables. To obtain a continuous variable that can be estimated, here it is opted to model the rankings via the *exploded logit* model which is also known as the *rank-ordered logit* or *Plackett-Luce* model (Beggs et al., 1981; Turner et al., 2018). This model allows ranked preference data and differs from the usual logit model of qualitative choice as it considers the ratings of all alternatives. The usual logit model only considers the most preferred alternative during modelling. By acknowledging all the information from the rankings, a more precise estimate of the *worth* of each alternative can be computed.

The worth of the fragments is obtained with the R package PackettLuce (Turner et al., 2018). According to Turner et al. the model is based on Luce's axiom which describes that for a given set S of J items, the probability of selecting an item j with worth $\quad_i$ for item i can be deduced with:

$$P(j|S) = \frac{\sum_{i \in J} i_j}{\sum_{i \in J} i} \tag{3.1}$$

Thus the probability of picking j is the worth of said item divided by the sum of the worth of all items in set S. Thereafter, the ranking can be seen as a sequence of choices where the first choice is the item ranked the highest and thereafter the second choice is the highest ranked item of the remaining set (Turner et al., 2018). This leads to the following model M which was derived by Plackett (1975):

$$M = \prod_{j=1} \frac{\sum_{i \in J} i_j}{\sum_{i \in J} i} \tag{3.2}$$

Thus the worth of an item is computed by multiplying the probability of selecting the item from several sets of alternatives. The higher the probability is that an item will be chosen in multiple different sets of alternatives, the higher the worth. Therefore, the different rankings as given in the user study are obtained and counted on how often they occur and in which sets of alternatives. With these rankings the Plackett-Luce worth is computed. To allow negative values, the log of the worth is taken, and thereafter the values are scaled to have a standardised continuous worth value to indicate the representativeness of fragments.

3.4.4 Familiarity

During the user study, participants are also asked to indicate whether they are familiar with the songs. To see whether familiarity has an impact on the rankings of songs, a familiarity score is computed by dividing the amount of respondents that were familiar with the song with the amount of unfamiliar respondents. As there were songs in the corpus which were either familiar or unfamiliar for every participant, plus one was added to both values:

$$Familiarity = \frac{Known + 1}{Unknown + 1} \tag{3.3}$$

To ensure that the familiarity values do have negative values, the log is taken from Equation 3.3. The last step is to scale the values to ensure compatibility and easier interpretation in comparison with the other features (the factors as well as the Plackett-Luce worth) which are also standardised:

$$Familiarity = standardised\ log\ \frac{Known + 1}{Unknown + 1} \tag{3.4}$$

3.5 Worth Approximation

To see whether the factors based on the CATCHY features as well as familiarity can explain the representative worth of fragments, a Generalised Linear Model (GLM) is used. A GLM is a model which uses a linear combination of independent variables to explain a non-linear dependent variable (Bishop, 2006; Nelder and Wedderburn, 1972). This means that similarly to linear models a linear function is created in the simplest form of

$$y(x; w) = w_0 + x_1 w_1 + ::: + x_n w_n \tag{3.5}$$

where $x = (x_1; :::; x_n)$ is the set of independent features and the weights $w_0; :::; w_n$ are the intercept and coefficients that are computed. To allow for linear combinations of non-linear functions, this definition of linear models can be extended to

$$y(x; w) = w_0 + \sum_{j=1}^{M-1} w_j \ _j(x) \tag{3.6}$$

where $_j(x)$ are basis functions and M as the amount of the total number weights. The addition of the GLM to the linear model is that the dependent variable does not need to have a normal distribution (Nelder and Wedderburn, 1972). However, as the dependent variable in this study is standardised, the GLM in this study is the same as a normal linear model.

Thus, for the approximation the obtained factors and the familiarity scores are used as independent values for the model. The GLM then tries to find a linear combination of these features to predict the dependent variable, the worth of the fragments. Features with high estimates that have a low probability to occur are most likely to be able to predict how representative each fragment is and can thereafter be used for the proposed thumbnailing method.

Additionaly, a second GLM is created which also considers the segmentation method as a categorical variable. Previous catchiness studies have indicated that sectional boundaries improve recognition, which could also mean that the segmentation method might have impacted the worth of segments.

3.6 Thumbnail Selection

Finally, the last part of this study is to create a new music thumbnailing method based on the findings. The results of the GLM show which factor or factors contribute to the explanation of the worth of a fragment significantly. These are used to create a "catchy" function which can assign a representative worth to new fragments. This function is a combination of the intercept and the estimates for the significant factors along with a computation to simplify these relevant factors.

The simplification of the factor scores is based on the features with high loadings (above 0.4) for the relevant factors. Having high loadings means that these features have a high contribution to the factor scores. For each of the important factors, new GLMs are run with the features with high loadings as independent variables. This should results in a simplified version of the factor based on the contribution of the most important CATCHY features. The reduction of features to approach the factor scores should make it easier to compute as only a subset of the CATCHY features is needed.

The functions for the factor approximations can be combined with the "catchy" function in a form of summary prediction function. This function predicts a representative worth of a given fragment for which the CATCHY features were computed. The fragment of a song with the highest predicted worth can thereafter be used as music thumbnail.

Thus, the proposed music thumbnailing method would start with the segmentation of song to be evaluated. Then, the relevant CATCHY features are computed for each fragment. This is followed by computing the approximated worth and selecting the fragment with the highest worth as music thumbnail.

Chapter 4

Results

This Chapter discusses the results obtained via the methods described in Chapter 3. The first steps of the method about data, segmentation, and the user study do not yield specific results. However, checks are made to gain insight into the Muziekweb user demographics and into the responses recorded via Qualtrics. Thereafter, the several steps for dimensionality reduction and their results are shown. These results lead to the approximation of the user study results with the obtained factors via EFA, the familiarity score, and the segmentation method. Finally, the results leading to a proposed music thumbnailing method are given.

4.1 Muziekweb Users

One of the assumptions for the user study is that Muziekweb users do not differ from the general population. Figure 4.1 displays some basic demographic content Muziekweb obtained via their website. These statistics were obtained by Google Analytics based on cookies. The percentages in the right upper corner display how many of the visitors of the Muziekweb site had cookies to obtain this information. It can be seen that the gender of these visitors is fairly equal and also that the age of the visitors is fairly equally distributed over different age classes. However, Muziekweb does seem to have relatively few visitors in the age group between 18 and 24 years of age.

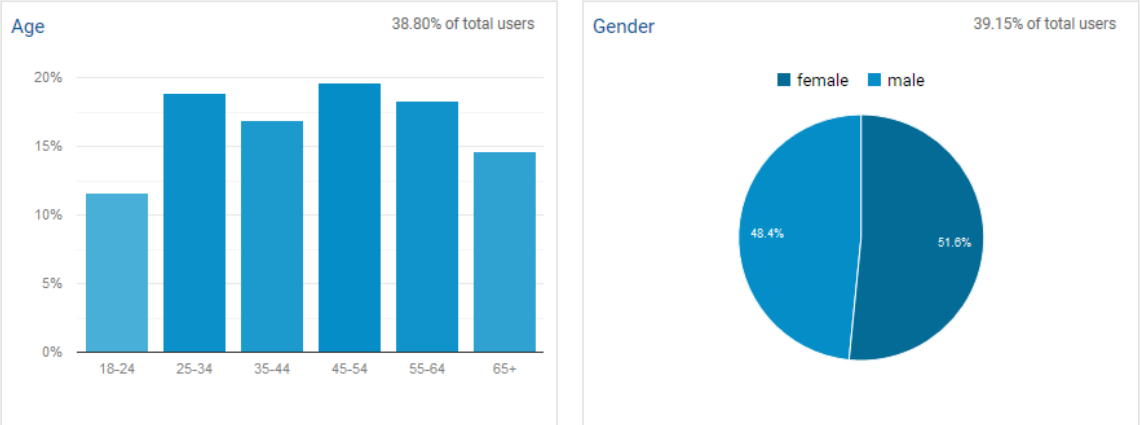

Figure 4.1: Demographics of visitors of the Muziekweb site displaying age and gender

4.2 Qualtrics Responses

The Qualtrics survey was available from April the 3rd, 2019 up to May the 27th, 2019 via a link posted on the site, Twitter, and Facebook of Muziekweb. Muziekweb made a post on April the 3rd, 2019 and a second time on April 23rd, 2019 to encourage respondents. At the same time, one of the researchers also posted the link on their Facebook. In this time frame, 148 responses were recorded, of which fourteen have completed the full survey and 93 have done so partially. The remaining responses did not get past the informed consent. The expectation was that participants would rank approximately 20 songs each. The mean amount of ranked songs was 25.35 (standard deviation of 20.81) and the median was 19.50. This was computed only for the 72 participants who managed to get past the example questions and shows that when participants were willing to actually participate, they would indeed rank approximately 20 songs.

This resulted in each segment being ranked with a mean of 15.25 (median of 15, standard deviation of 3.19). This displays only how often each segment was seen and ranked, and does not give an indication of how often each pair of segments was ranked. This is lower than the goal of 30 measurements per pair. However, the results later in this Chapter show that the amount of collected data was sufficient.

4.3 Dimensionality Reduction

Here is described how the dimensionality of the feature space was reduced, which steps and choices were considered, and what contributes to the obtained factors.

4.3.1 Correlation

The first step for dimensionality reduction is to create a correlation or covariance matrix based on the features. The type of the matrix depends on whether the feature values are normally distributed. To test this, the distributions of all obtained feature values with the CATCHY toolbox are plotted in Figure 4.2 which show this is not the case. Also notable are the bumps at the ends of some of the plots, which is caused by how the second-order features are computed for the one-dimensional features. However, these should not influence whether a feature is normally distributed. As the plots indicate that transformations may be needed to ensure that all features are normally distributed, it was opted to use the Spearman correlation as this correlation does not assume normal distribution of features. The resulting correlation matrix between features is hereafter used for the following dimensionality reduction steps.

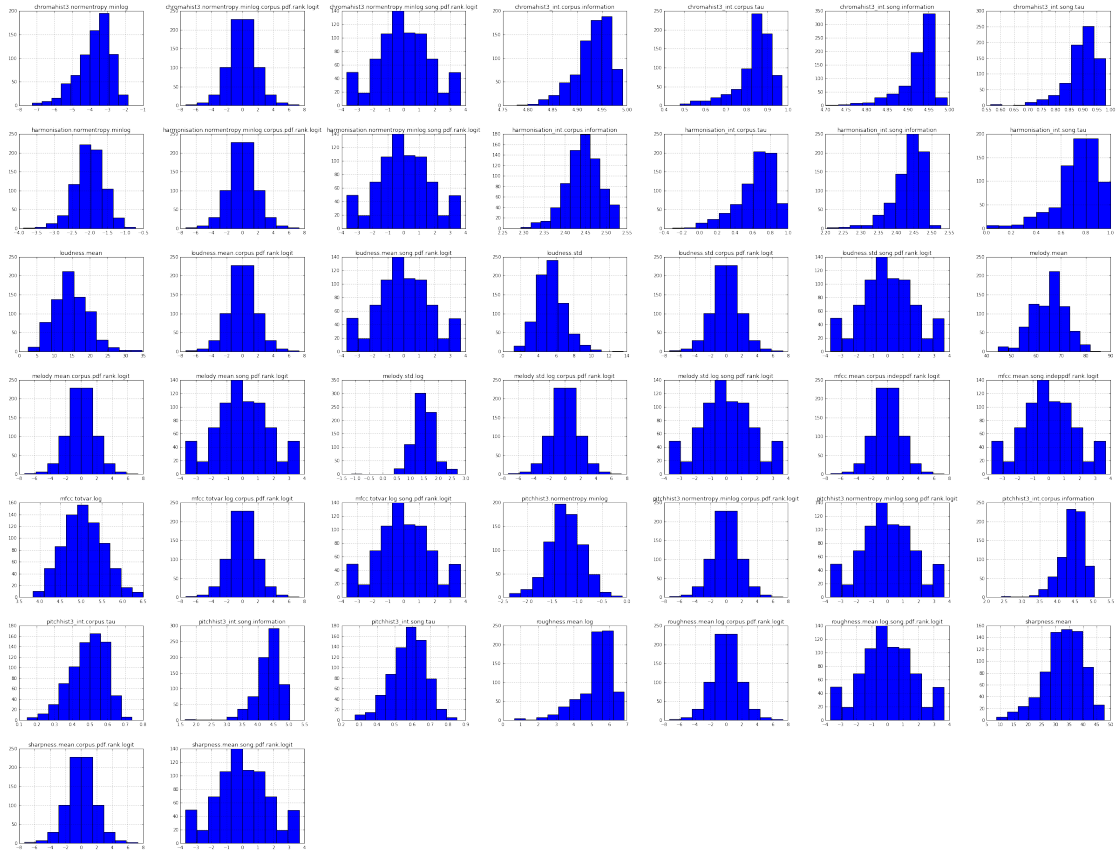

Figure 4.2: Distribution of the occurrences of all features computed with the CATCHY toolbox

4.3.2 Number of factors

The second step is to check how many factors, or dimensions, are needed to explain the variance in the data well enough. Figure 4.3 shows four plots of different methods to identify the best amount of factors. These plots were optimised for the usage of Varimax rotation and minimum residual factor analysis.

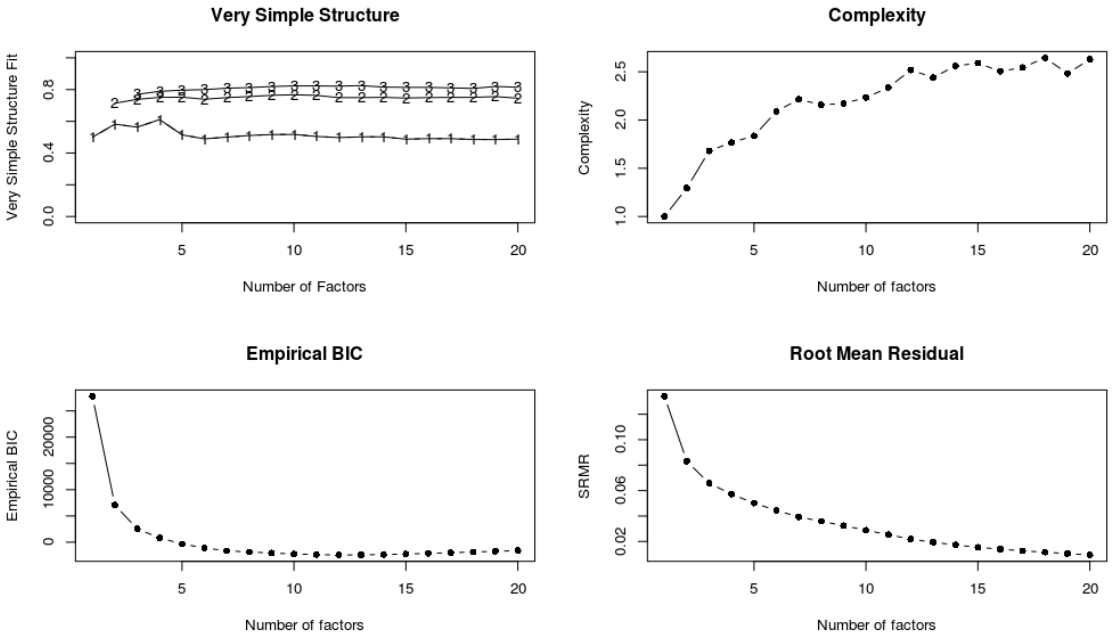

Figure 4.3: Results of different methods to find the optimal amount of factors

The figure shows no single obvious number of factors to use. The Very Simple Structure technique shows a small bump at four factors but is mostly stable. The complexity only seems to rise when factors are added, which also advocates for choosing less factors than the number of twelve chosen by Van Balen et al. (2015a). The Empirical BIC does show a lowest value with more factors but is relatively stable after reaching five or six factors. Lastly, the Root Mean Residual should in an ideal case show a clear elbow point, which is not the case here. Overall, the plots seem to indicate that there is no optimal number of factors, but keeping the number low does seem to be preferred. This is supported by also running IClust which showed that features were preferably clustered into three clusters (see Appendix D). This clustering does show one much larger cluster which could be divisible, indicating that three dimensions might not be enough to explain the data.

With these findings on the amount of dimensions that would work the best in mind, the results for EFA with the amount of factors between four and eight were tested and compared (see Appendix E for the loading tables). Models with more than six factors include factors with only two high loadings (at least a loading 0.4 or higher), which is an indication that six factor already has a too high dimensionality. Morover, models with more factors than five would have factors that would overlap too much. This is another indication of too many factors. While overfactoring is less harmful than underfactoring, it is key to choose as many factors without losing interpretability or identifiability. This lead to choosing between four and five factors. Here, five factors was preferred as the interpretability of the factors containing less high loadings is still possible, while they do contain at least three high loadings or more. Moreover, the four-factor model has more features that do not have high loadings in any of the factors, which emphasises that a five-factor model explains more of the variance in the data.

4.3.3 Factor Analysis with Five Factors

Table 4.1 shows the loadings of the different features for the minimum residual EFA model run for five factors with Varimax rotation. The table shows for each feature its loadings for the different factors (where MR stands for minimum residual), its communality (h2), its uniqueness (u2), and its complexity (com) of the component loadings (Revelle, 2011). The factor loadings show the importance of the features for the different factors. The communality measures how much of the variance of a feature is explained by all factors together. The uniqueness is measured by the variability minus the communality.

4.4 Representativeness Approximation

The obtained factors via the EFA are used as new features to approximate the user study rankings. Therefore, the factor scores based on these factors are computed for each segment of the user study. Thereafter, a GLM is created with as dependent variable the standardised worth of the fragments and the five factors and the familiarity scores as independent features. The results are shown in Table 4.2, which shows the coefficients (b) of the features, the standard error, as well as the t-score and p-value to indicate how likely the results were.

The results of the GLM indicate that MR3, MR4, and MR5 could explain the worth of the fragments significantly. To further test this assumption, the function

	Factors							
Feature	MR1	MR2	MR3	MR4	MR5	h2	u2	com
HIC Entropy	**-0.89**	-0.28	-0.02	-0.03	-0.04	0.86	0.14	1.21
HI *j* Corpus Information	**0.88**	**0.32**	0.06	-0.05	0.00	0.88	0.12	1.27
HIC *j* Corpus Information	**0.88**	0.30	0.07	0.02	0.04	0.87	0.13	1.25
HI *j* Song Information	**0.87**	**0.34**	0.06	-0.04	-0.01	0.89	0.11	1.32
HIC *j* Song Information	**0.86**	**0.31**	0.07	0.02	0.03	0.83	0.17	1.27
HI *j* Corpus	**-0.61**	0.08	0.03	0.20	0.05	0.42	0.58	1.27
HI *j* Song	**-0.52**	0.08	0.06	0.23	0.14	0.36	0.64	1.63
HIC *j* Corpus	**-0.50**	0.11	0.18	0.21	0.18	0.37	0.63	2.08
HIC Entropy *j* Corpus	**-0.48**	-0.02	0.24	0.16	0.16	0.34	0.66	1.97
HIC *j* Song	**-0.45**	0.05	0.24	0.17	0.19	0.33	0.67	2.32
Melodic Pitch SD	0.08	0.00	-0.04	0.03	-0.06	0.01	0.99	2.82
MIB Entropy	**-0.37**	**-0.90**	0.00	-0.02	-0.06	0.95	0.05	1.35
MIB *j* Song Information	**0.38**	**0.87**	-0.03	0.02	0.03	0.91	0.09	1.38
MIB *j* Corpus Information	**0.43**	**0.84**	-0.03	0.01	0.05	0.89	0.11	1.51
HI Entropy	**-0.50**	**-0.76**	-0.08	0.00	-0.05	0.84	0.16	1.77
Melodic Pitch SD *j* Corpus	-0.06	**0.33**	0.20	0.00	-0.03	0.16	0.84	1.75
MIB *j* Corpus	-0.26	**0.32**	0.08	0.30	0.14	0.28	0.72	3.48
Melodic Pitch SD *j* Song	-0.08	0.17	0.15	0.02	0.00	0.06	0.94	2.48
Loudness	0.04	-0.06	**0.93**	0.03	-0.04	0.87	0.13	1.02
Roughness	0.17	0.08	**0.79**	0.09	0.14	0.68	0.32	1.21
MFCC Variance	0.24	0.03	**-0.54**	0.01	-0.04	0.35	0.65	1.39
Loudness SD	0.30	0.06	**0.45**	0.14	0.04	0.32	0.68	2.01
MFCC Mean *j* Corpus	-0.03	0.21	**0.45**	0.20	**0.35**	0.41	0.59	2.84
Melodic Pitch Height	0.17	-0.06	**0.45**	0.08	-0.01	0.24	0.76	1.41
Sharpness *j* Corpus	0.12	0.13	**0.32**	0.15	0.20	0.19	0.81	2.91
MIB Entropy *j* Corpus	-0.08	0.04	0.12	**0.77**	-0.03	0.61	0.39	1.08
HI Entropy *j* Corpus	-0.10	0.06	0.09	**0.76**	0.01	0.60	0.40	1.07
HI Entropy *j* Song	0.04	-0.08	0.00	**0.52**	0.19	0.32	0.68	1.33
MIB Entropy *j* Song	0.02	-0.12	0.00	**0.50**	0.09	0.27	0.73	1.19
MIB *j* Song	-0.09	0.28	0.05	**0.32**	0.17	0.22	0.78	2.74
Loudness *j* Corpus	0.03	0.18	-0.14	-0.02	**0.56**	0.37	0.63	1.35
Roughness *j* Corpus	0.06	0.06	**0.41**	0.03	**0.52**	0.44	0.56	1.97
Roughness *j* Song	-0.04	0.07	0.25	0.03	**0.48**	0.30	0.70	1.59
Loudness *j* Song	-0.03	0.09	-0.02	0.06	**0.46**	0.22	0.78	1.13
Sharpness	0.20	**0.35**	0.18	0.18	**0.43**	0.41	0.59	3.24
Loudness SD *j* Corpus	-0.02	0.15	0.03	0.02	**0.34**	0.14	0.86	1.44
MFCC Mean *j* Song	-0.04	0.09	0.19	0.16	**0.31**	0.17	0.83	2.46
MFCC Variance *j* Song	-0.03	0.07	0.10	0.05	0.28	0.10	0.90	1.49
Melodic Pitch Height *j* Corpus	0.05	0.08	0.19	0.15	0.27	0.14	0.86	2.76
Loudness SD *j* Song	0.00	0.10	-0.07	0.00	0.25	0.08	0.92	1.51
HIC Entropy *j* Song	-0.06	0.00	0.22	0.06	0.25	0.11	0.89	2.23
MFCC Variance *j* Corpus	-0.08	0.14	0.00	-0.05	0.23	0.08	0.92	2.07
Melodic Pitch Height *j* Song	0.00	-0.01	0.16	0.04	0.20	0.07	0.93	1.98
Sharpness *j* song	0.04	0.12	0.14	0.06	0.16	0.06	0.94	3.27
SS loadings	6.28	4.03	3.22	2.27	2.22			

Table 4.1: Factor loadings for Minimum Residual EFA with five factors showing the factor loadings, communalities (h2), uniquenesses (u2), and complexities of the factor loadings (com)

"dredge" from the MuMIn package in R is used which creates models with all possible combination of independent variables to explain the dependent variable. Table 4.3 shows the results of the nine best performing models. Models could be interchangeably used when their ΔAIC_c is lower than three. However, as the factors as well as the familiarity score are standardised, it should be noted that the size of the coefficients matters. Thus, even when the ΔAIC_c suggest models which start to include MR2 or familiarity can also be used, the coefficients of these variables are so small that they do not add any value.

Feature	b	SE	t	p
Intercept	-0.04	0.05	-0.75	0.452
Familiarity	-0.08	0.07	-1.11	0.267
MR1	-0.05	0.05	-0.90	0.370
MR2	0.06	0.05	1.19	0.236
MR3	0.17	0.05	3.18	0.002
MR4	0.20	0.05	3.91	<0.001
MR5	0.30	0.05	5.64	<0.001

Table 4.2: GLM results using factors and familiarity. For each variable, the estimate or coefficient (b), the standard error (SE), the t-score, and p-value are given

Intercept	Familiarity	MR1	MR2	MR3	MR4	MR5	df	ΔAIC_c
-0.05				0.16	0.20	0.29	5	0.00
-0.05			0.06	0.17	0.20	0.29	6	0.64
-0.04	-0.07			0.16	0.20	0.30	6	1.00
-0.05		-0.04		0.17	0.20	0.29	6	1.29
-0.04	-0.07		0.06	0.16	0.21	0.30	7	1.58
-0.05		-0.04	0.06	0.17	0.20	0.29	7	2.02
-0.04	-0.08	-0.05		0.17	0.20	0.30	7	2.19
-0.04	-0.08	-0.05	0.06	0.17	0.20	0.30	8	2.86
-0.05					0.20	0.30	4	7.73

Table 4.3: The best nine models to approximate the representative worth based on different combinations of independent features. Along with the estimates for each feature, the degrees of freedom (df) as well as a measure to compare models (ΔAIC_c) are shown

Feature	b	SE	*t*	*p*
Intercept	-0.03	0.13	-0.20	0.838
familiarity	-0.08	0.07	-1.10	0.275
MSAF	-0.10	0.14	-0.72	0.474
Random	0.33	0.18	1.84	0.067
MR1	-0.05	0.05	-0.91	0.365
MR2	0.05	0.05	1.03	0.306
MR3	0.16	0.05	3.08	0.002
MR4	0.22	0.05	4.20	<0.001
MR5	0.30	0.05	5.41	<0.001

Table 4.4: GLM results with the addition of a categorical variable for the method of segmentation with as standard the 1-minute in method

Feature	Chisq	Df	*p*
Familiarity	1.20	1	0.275
Segmentation	9.44	2	0.009
MR1	0.82	1	0.364
MR2	1.05	1	0.305
MR3	9.50	1	0.002
MR4	17.63	1	<0.001
MR5	29.25	1	<0.001

Table 4.5: ANOVA results to test the importance of the segmentation method

One hypothesis during the set-up of this study was that hooks lie at the beginning of structural sections. Therefore, a second GLM was created with the addition of how the fragments were obtained as categorical variable (either via MSAF, random selection, or 1-minute-in). The GLM with this additional categorical variable does indicate that the segmentation method could affect the score as can be seen in Table 4.4. Based on these results, a post-hoc comparison is carried out to check the influence of the categorical variable as a whole in the form of an ANOVA from the "car" package in R (see Table 4.5).

The ANOVA indicates that the mean worth per segmentation method may differ significantly. An additional post-hoc test "glht" is run via the "multcomp" R-package[a] which performs a Tukey test to compare the segmentation groups (Hothorn et al., 2016). The results of this test are extracted via "cld" which creates a compact letter display of all pair-wise comparisons along with a box plot showing the distributions of the dependent value for each segmentation method.

[a] https://cran.r-project.org/web/packages/multcomp/index.html

The results are shown in Figure 4.4 where the strict interpretation of the letter plots is that for any two groups that share a letter, there is insufficient information to reject the null hypothesis that the groups performance is the same. In this case, this means that the 1-minute-in segmentation method does not differ from random or MSAF segmentation. However, as random and MSAF segmentation do not share a letter, they are different according to this test.

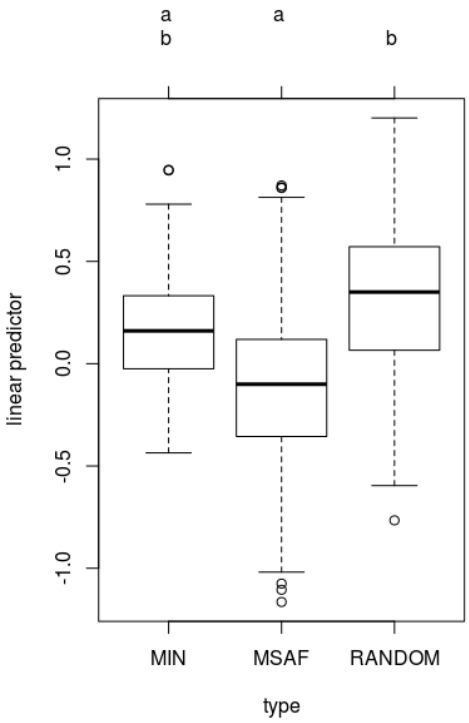

Figure 4.4: Plot of the worth per segmentation type with the letters defining which groups are similar

4.5 Catchy Function

Lastly, the proposed music thumbnailing method based on the results can be established. The results of the GLMs show the importance of the factors MR3, MR4, and MR5 to approximate the worth of an fragment. However, the factors rely on the full set of 44 features in the CATCHY toolbox. Therefore, to simplify the derivation of the feature scores, new GLMs are created for each of these three factors using the factor scores as dependent variables and the features with high loadings (loadings above 0.4) as independent variables. The results of these models are shown in Table 4.6, 4.7, and 4.8 for factors MR3, MR4, and MR5 respectively. For each factor, the scores can be approximated by a summation of the features with a significant p-value multiplied by the estimates of the respective GLM. Thereafter, these approximations are combined with the results of the first model in Table 4.3 to create a function for representativeness. These estimates are preferred as they are computed without the

influence of familiarity, MR1, nor MR2. This combination results in Equation 4.1 which computes the representative worth of a given fragment. However, as second-order features are based on the reference corpus used, there are some risks in using them as they change depending on the music corpus and segmentation.

$$
\begin{aligned}
\mathit{worth} = \quad & 0{:}05 + 0{:}16 \quad \mathit{MR3} + 0{:}20 \quad \mathit{MR4} + 0{:}29 \quad \mathit{MR5} \\
= \quad & 0{:}620 + 0{:}024 \quad \mathit{Loudness} + 0{:}013 \quad \mathit{Roughness} \\
& 0{:}021 \quad \mathit{MFCC\ Variance} + 0{:}013 \quad \mathit{MFCC\ Mean\ j\ Corpus} \\
& + 0{:}034 \quad \mathit{MIB\ Entropy\ j\ Corpus} \ + 0{:}018 \quad \mathit{MIB\ Entropy\ j\ Song} \\
& + 0{:}032 \quad \mathit{HI\ Entropy\ j\ Corpus} + 0{:}018 \quad \mathit{HI\ Entropy\ j\ Song} \\
& + 0{:}060 \quad \mathit{Loudness\ j\ Corpus} + 0{:}039 \quad \mathit{Loudness\ j\ Song} \\
& + 0{:}030 \quad \mathit{Roughness\ j\ Corpus} + 0{:}039 \quad \mathit{Roughness\ j\ Song} \\
& + 0{:}006 \quad \mathit{Sharpness}
\end{aligned}
\tag{4.1}
$$

Feature	b	SE	*t*	*p*
Intercept	-2.36	0.31	-7.55	<0.001
Loudness	0.15	0.01	21.63	<0.001
Roughness	0.08	0.03	2.48	0.013
MFCC Variance	-0.13	0.05	-2.78	0.006
Loudness SD	0.00	0.01	0.31	0.755
MFCC Mean *j* Corpus	0.08	0.01	7.08	<0.001
Melodic Pitch Height	0.01	0.00	1.87	0.062

Table 4.6: Approximation of MR3 by GLM

Feature	b	SE	*t*	*p*
Intercept	0.00	0.00	0.00	1
MIB Entropy *j* Corpus	0.17	0.02	7.73	<0.001
MIB Entropy *j* Song	0.09	0.02	4.44	<0.001
HI Entropy *j* Corpus	0.16	0.02	7.38	<0.001
HI Entropy *j* Song	0.09	0.02	4.31	<0.001

Table 4.7: Approximation of MR4 by GLM

Feature	b	SE	t	p
Intercept	-0.64	0.13	-4.79	<0.001
Loudness j Corpus	0.20	0.02	12.60	<0.001
Loudness j Song	0.13	0.02	8.01	<0.001
Roughness j Corpus	0.10	0.02	6.11	<0.001
Roughness j Song	0.13	0.02	7.75	<0.001
Sharpness	0.02	0.00	4.87	<0.001

Table 4.8: Approximation of MR5 by GLM

Chapter 5

Discussion

Whether the notion of catchiness can be used for music thumbnailing is dependent on several steps in this study. First the factors from the EFA are interpreted and evaluated. Then, the results of the GLMs are considered to assess the importance of the factors as well as familiarity and segmentation method. Thereafter, the catchy function and how this would be used for music thumbnailing is discussed. This is followed by limitations of this study and further research.

5.1 Analysis

The analysis and interpretation of the results in the several steps is important to see what they could mean and whether the research question is answered. First an interpretation of the EFA factors is derived to know what each factor is possibly measuring in the audio signals. Then, the PCA components are interpreted and compared to the factors and the components found in Van Balen et al. (2015a). This should give an idea of how the choice for EFA might have impacted the results. This is followed by the several GLMs that were trained to see whether the inputted independent variable could explain the user study rankings. Then, the proposed method for Muziekweb is shown with preliminary results on the effectiveness of the thumbnailing method. Lastly, the current study is compared to previous studies.

5.1.1 Factor Analysis

The first part of analysing the results is by interpreting what the obtained factors via the EFA could measure in the music as shown in Table 4.1. Here, an interpretation

of these factors is given by looking mostly at the features with high loadings of above 0.4.

The first factor, MR1, consists mostly of second-order higher-dimensional features describing harmonics. Both the HI which shows the harmonisation between the melody and the chords, as well as the HIC which describes triads have very high and similar positive loadings for their second-order feature measured in terms of information-theoretic unexpectedness, or Information. The second-order features for HI and HIC computed via Kendall's also have similar loadings but add negatively to this factor. This could mean that all these different measures for the higher-dimensional features do measure the same thing and thus describe the same variance in the data. A higher Information score means that the feature value has a lower probability to occur; there is more information needed to describe the event. This aligns with a lower as this indicates that there were less occurrences of similar values. The remaining high loadings in this factor are the first-order measure of the HIC and HI in terms of entropy and second-order features of the HIC based on entropy. For the first-order HIC and HI, entropy is used as a first-order measure of dispersion which means that a lower value represents less dispersed data (Van Balen et al., 2015a). This factor thus prefers less dispersed HIC and HI as can be seen by the high negative loadings. Furthermore, the second-order entropy features for HIC could also indicate that a less common dispersion of data is preferred. Lastly, features based on MIB also have some higher loadings, with one above the threshold of 0.4. This shows that apart from harmonic uniqueness, the factor also likely measures melodic uniqueness to some degree. This most likely means that this factor measures the uniqueness of values for the different higher-order harmonic (and also slightly melodic) features in relation to different corpora.

Factor MR2 is very similar to MR1 but has very high loadings for the melodic higher-dimensional features. Similarly, the second-order features computed via information add positively to this factor, while the entropy for MIB and HI adds negatively. There is also quite some overlap between this factor and the first one in loadings above 0.3, meaning that this factor might emphasise the melodic higher-dimensional features but is still very similar to the harmonic higher-dimensional features. This overlap could suggest that a four-factor model would have been preferred. However, even in a four-factor model, there is still a lot of overlap between factors depending on second-order higher-dimensional features (see Appendix E). At the same time, more features in the four-factor model have a low communality (h2) which indicates that they are not represented enough. Despite the overlap of this factor with MR2, it is thus preferred to explain more variance with a five-factor model.

Thereafter, the factors start containing the basis features and their second-order

variant. Simultaneously, they start including smaller sets of features, starting with MR3. This factor mostly relies on a high positive values for loudness (mean and standard deviation) and roughness. It also prefers a lower MFCC variance, a high MFCC mean, and a high melodic pitch height. This factor should thus measure the intensity of fragments.

The fourth factor, MR4, is heavily based on the corpus as well as song-based second-order features for the MIB and HI entropy. This means that the values for this factor rise when the dispersion of MIB and HI is more common. Thus, it is a measure that describes the commonality and recurrence of melodic and harmonic aspects in comparison with the reference corpus. This could thus both indicate repetition within a song itself, while also maintaining sounds that do not stray too far from the corpus.

Lastly, factor MR5 comprises of less high loadings in comparison with the other factors, with only two features that score above 0.5 and two that almost near it. These four features describe the corpus and song-based second-order features for loudness and roughness with a preference for the corpus-based variants. The last feature with a high loading is sharpness. This factor thus probably describes the commonness of the intensity (as the features overlap with those in factor MR3) in relation to the full corpus and the song. Therefore, this factor depicts the commonality and recurrence of the intensity of the fragment in combination with a high sharpness.

These interpretations show that the factors describe higher-dimensional harmonic and melodic uniqueness, intensity, and recurrence and commonality of fragments.

5.1.2 Generalised Linear Model and Implications

The next step is to evaluate the GLM made with the interpreted factors and the familiarity score as independent variables. The results of this model is shown in Table 4.2. The factors MR3, MR4, and MR5 have high estimates (b) in this model along with significant p-values. As all variables in the model are standardised, the size of the coefficients matters. This also means that the other two factors and the familiarity score are unimportant for the approximation of the worth of a segment. The importance of the three factors is strengthened by the results shown in Table 4.3. These results indicate that the model with the least degrees of freedom (df), and thus the least complex but working model, is a model only consisting of these three factors as features. The relative quality of the models is given by the ΔAIC_c and shows that most tested models can be used interchangeably as the values stays below three. The table thus shows the first model lacking one of these three factors has a great decline of the ΔAIC_c, indicating the importance of this factor for the

model. Moreover, the estimates of the MR3, MR4, and MR5 remain consistent in each set-up of features, while the estimates for the other independent features remain low. This shows that the factors MR3, MR4, and MR5 are the best indicators of the dependent variable.

With the interpretation of the factors in mind, it is possible to infer what these factors are and how they contribute to better representativeness as indicated by the Plackett-Luce worth. The most important contributor with an estimate 0f 0.30 is MR5, which thus means that if the intensity of the fragment is more common in comparison to the corpus, but also in the song, the worth rises. MR4 has an estimate of 0.20 and describes whether the melodic content as well as the harmonisation between melody and chords has a dispersion that is common in the reference corpus. These two factors and their importance thus indicate that a more representative fragment has a higher commonality and recurrence. Lastly, MR3 has the lowest significant estimate of 0.17, which means that the representativeness of a fragment increases when the intensity of the fragment is higher.

Combining these interpretations, fragments that are more representative are most likely fragments that have high recurrence. This could mean that they are often repeated within a song. They should also not stray from the full corpus, indicating that they have a generally accepted sound within the corpus and thus a high commonality. Lastly, the intensity of the fragment should be high. This means that within popular music the most representative fragments could be the fragments containing the chorus as those are often the most repeated parts of a song with a relatively high intensity (Bartsch and Wakefield, 2005; Huang et al., 2017).

Apart from testing the features of the CATCHY toolbox, familiarity was also evaluated with the same GLM on whether it has an effect on representativeness. In this model, the estimate for the standardised familiarity score is rather low and seems to indicate that familiarity could influence the worth of a fragment negatively. However, with a very low estimate and insignificant p-value, it is safe to say that familiarity does not impact the worth of a fragment. Thus, listeners are able to evaluate the representativeness of a fragments without the need of knowing the song.

Lastly, an assumption made in this study was that fragments starting at the beginnings of structural sections might have a higher chance of containing the most representative part of the song. This was tested by running another GLM with the addition of the segmentation method as categorical variable. The results in Table 4.4 show that the segmentation method may have an influence, with a preference for random segmentation. Therefore, an ANOVA was run as on the model as post-hoc test which shows that segmentation does seem to have a significant effect on the worth (see Table 4.5). To see how the different segmentation methods differ, Figure

4.4 shows the results of a pair-wise comparison of the different segmentation groups. The compact letter display shows that segmentation by taking a fragment starting at 1-minute in does not differ from fragments segmented via MSAF nor random selection. However, segmentation via MSAF does differ significantly enough from randomisation, where segments that are randomly selected have a better worth in this data set.

These results shows that the hypothesis of segments starting at structural beginnings would benefit representativeness is not supported. This could be due to the aim of the study. Where Burgoyne et al. (2013) tried to identify the exact moment in a song were recognition starts, here the focus is on creating thumbnails. A thumbnail does not necessarily improve by starting at a point of recognition, but benefits the most of containing the most representative part somewhere in the thumbnail. This does not mean the most recognisable part needs to be at the exact beginning of the thumbnail. Moreover, the set-up of the Hooked user study awarded participants to recognise a song as fast as possible, meaning that the start of the fragment is much more important than in the current set-up.

5.1.3 Catchy Thumbnailing Function

With these results in mind, a proposition for a automatic music thumbnailing method is made. This is mostly based on the GLM results indicating the importance of MR3, MR4, and MR5 with estimates of 0.17, 0.20, and 0.30 respectively.

The first part of the method would be to use a segmentation method to obtain different fragments of the same song. The results show that no specific segmentation method is preferred. Then, for each of these fragments an approximation of the representative worth can be computed by computing the CATCHY features and inserting the values into Equation 4.1. It should be noted that because these factors are computed with second-order features which depend on their respective reference corpus, the results will change.

The second step would be to choose a fragment based on the computed worth for all fragments created with the segmentation method. The best method would be to choose the fragment with the highest worth and select this as chosen thumbnail for the song. Th evaluation of this proposed method would be part of a follow-up study.

5.1.4 Comparison to Other Studies

The previous subsection describes the proposed method for music thumbnailing based on the results from this study. The interpretation of the factors that contribute

the most to this method indicate that intensity, and a higher commonality and recurrence are related to the representativesness of a fragments. This aligns with previous automatic thumbnailing studies, as they mostly focused on detecting the most repeated section or chorus (Bartsch and Wakefield, 2005; Chai and Vercoe, 2003; Huang et al., 2017; Levy and Sandler, 2006; Müller et al., 2013; Schuller et al., 2008). This means that the results based on the preferences of listeners strengthen the assumptions of previous studies where the best thumbnail was thought to be the part of the music that is repeated most often, which often aligns with the chorus.

The importance of intensity, commonality and recurrence also aligns with previous work aimed at catchiness and INMI. Van Balen et al. (2015a) found that fragments that are easier to recall have a more typical sound, more conventional melodies, more recurrence in the timbral aspects, and a prominent vocal line. The last aspect could not be measured with the factors in this study. Moreover, INMI, which is related to catchiness, seems to appear more often for often repeated fragments with a faster tempo and a common melodic contour (Jakubowski et al., 2017; Williamson et al., 2012). The results of this study thus agree with previous works, meaning that the assumption that catchiness also overlaps with representativeness seems plausible.

5.2 Limitations

While many results are very significant, several choices were made that could have limited this study. The impact of the data, segmentation, user study format, as well as the creation and processing of the features and models are discussed.

5.2.1 Data

The first choice in this study is how data was obtained and chosen. The data that was used is based on the listening counts of visitors of Muziekweb's site. It was noted that there was a bias towards French and German songs likely caused by Google listing Muziekweb as the first option for these terms. This makes it possible that listening counts are based on casual visitors of the Muziekweb site as well. This limitation, however, is not within the reach of this study. Figure 4.1 shows that the demographics of Muziekweb users does not stray from the general population in terms of age or gender. While there can still be differences between the general population and Muziekweb users and visitors in terms of other criteria, the user study results should also be based on a mix of Muziekweb users and visitors, as well

as people falling outside of this scope. Regarding the significance of this study, it is unlikely to have affected the results.

Another choice concerns the selection of data. During selection, songs were ignored from the same artists and albums to increase the diversity in the corpus. An exception was made for compilation albums, which took up a large part of the list, with the assumption that compilation albums do contain a representative collection of popular music. While it still limits the broadness of the study, choosing a list based on Muziekweb's listening counts was to ensure that participant would be familiar with the music. As the results have shown that familiarity is not necessary for representativeness, the importance of familiarity could already be measured.

The last limitation of the data selection is that only pop music is included. The results in this study are significant, but no statements can be made about other musical genres as thumbnail preferences differ between genres (Schuller et al., 2008).

5.2.2 Segmentation

The results in this study showed that fragments do not need to start at structural boundaries to be more representative. There was even a slight preference for segmentation by choosing a random starting point instead of a boundary chosen by MSAF. A part of this observation might be caused by randomisation involved in the selection of fragments chosen by MSAF. The MSAF algorithm normally detected at least four or more boundaries, which exceeds the amount of fragments that were used in the user study. This randomisation of choosing detected boundaries arbitrarily might have had a negative effect on the chosen segments. Thus, complete random selection of the detected boundaries might have left the best fragments of the song starting at structural boundaries out in the user study. However, this cannot completely explain the effect encountered in this study. A follow-up study could investigate the cases where boundary sections lack in comparison to randomly chosen starting points to investigate what happens in the music.

Additionally, only one boundary detection algorithm has been evaluated in the user study. This study thus only reveals that the implementation of the algorithm in MSAF based on Serra et al. (2014) is not necessarily preferred for the segmentation of representative fragments. This study thus does not give an indication on how other segmentation algorithms would affect the results. Another choice for the segmentation algorithm was also to choose on which type of time series the boundaries would be detected. While using chromagrams gave plausible results and aligns better with the features extracted via the CATCHY toolbox, it is a limitation on what aspects of the audio signal are considered during boundary detection. Further research

could look into the differences between boundary segmentation algorithms, which could give insight into whether boundary segmentation in general is not preferred for identifying representative fragments, or only some specific algorithms.

5.2.3 Qualtrics

Creating the user study in Qualtrics had several pros and cons. While the platform is relatively easy to use, data is stored securely, and playing music did not cause any problems, the choices that needed to be made might have influenced this study as well.

One such choice is to not add progress bars to the study. The idea behind this is that seeing the length of the full study might discourage participants to continue as their progress would be really slow. But, for participants who want to know what they are up to when participating in a study, might stop due to not being able to follow their progress. In this study, leaving the bar out still made more sense as it was not expected of participants to fulfil the complete study. Leaving the progress bar out should have encouraged participants to evaluate as many songs as they wanted to.

Another choice was to use the drag-and-drop option in Qualtrics for ranking. While this option is the most mobile-user friendly, it also has some limitations. One of them is that the response is not recorded when none of the alternatives are moved at least once. To ensure that everything is recorded, a warning is shown to the participant when they have not moved one of the fragments at least once before proceeding. This does solve the problem of recording responses, but causes the ranking to be more tedious than needed. Participants need to move fragments even when their preferred ordering is already given and if they forget to do so, end up with a warning message which can be annoying over time.

Another problem with the drag-and-drop function arises when music players are used as alternatives. The boxes cannot be moved by clicking on the players themselves, making it less instinctive. Moreover, the size of the music players cannot be adjusted, meaning that the boxes are too broad to fit the average mobile phone screen. This makes the survey drastically worse on mobile, while Muziekweb has indicated that half of their users visits their site via a mobile device. However, the option of using drag-and-drop ranking was still preferred above the other ranking option in Qualtrics where the limitations for mobile users would be greater due to typing or having to click on a ranking that does not fit the typical phone screen.

The design of the study might also have had an impact, such as to have an introduction followed by a page containing the informed consent. A large introduction

and an intimidating informed consent might have caused nearly one-third of the participants to discontinue the survey at the user study. While better phrasing might have made the introduction more compact, the intimidating informed consent is a must for a scientific experiment.

Lastly, there could also be an effect due to the exact phrasing used in the user study and the symbols used to distinguish between the music players. The significant results show that it is likely participants did have a similar grasp on what was expected from them to rank, even when "an idea of the song" is vague wording and could differ between participants. However, this phrasing was used to ensure participants would rank items based on a gut feeling without mulling on what representativeness might mean. Also, while individual participants might have had ranked preferences for the filled squares, there is no common ranking and thus, having enough participants would overcome personal preferences.

Thus, certain choices regarding the user study might have impacted how enjoyable it was for respondents to participate and might have caused bias. However, the clear results show that participants knew what they were supposed to do in the study and personal biases should have been overcome. Yet, future studies wanting to conduct a similar user study may consider a different platform for questionnaires.

5.2.4 Model and Analysis

The last impact by choices is in the processing of the data and modelling afterwards. First, the features that are used are only those currently available in the CATCHY toolbox. This means that rhythmic features were not considered in this study, while previous research has shown that melodies containing longer average durations are more likely to become INMI (Müllensiefen and Halpern, 2014). As INMI are related to catchiness, this could also be the case for representativess.A follow-up study could start by also considering the relationship between this branch of features could and representativeness.

Second, the features are not used themselves to explain the user study scores. Instead, they are used for a five-factor model created via EFA. Dimensionality reduction does help to describe variance in less dimensions and creates easier to interpret features. However, the procedure has many options and very few guidelines (Osborne et al., 2008). Here, the choice for an EFA is demonstrated as being profitable in comparison to a PCA as only shared variance is considered which makes it easier to interpret while an underlying structure is considered. However, more choices were made, such as doing an EFA via minimum residual and using Varimax rotations. Different choices here might have given different results. But, while for this study

different choices could have been made, the made choices did result in interpretable factors that were able to explain the representative scores via a GLM.

Third, the approximation of the representative worth of the fragments is done via a linear model. Although linear models are very easy to interpret, it is a large assumption to claim that a linear model would accurately model a complex world. Models allowing different relationships between independent and dependent variables might represent the actual relations better, but would also make the results far less clear. A follow-up step would be to also test more complex models and to compare these with the linear model. This could give insight into whether a linear model is able to sufficiently explain the ascribed representativeness scores.

Lastly, the results from this study are simplified for the proposed method. Table 4.3 does show that using only three factors for the approximation of the worth of a fragment is sufficient. Thereafter, a simplifcation of the factor scores is proposed. However, whether the use of this limited set of features is enough to represent the factors is not tested here. It could be that using the actual factors for the computation of worth may be preferred as the factor scores are more accurate even though the computation of features becomes more arduous.

5.3 Further Research

The discussion of limitations already discussed several considerations for further research. These were possible follow-up studies which consider different genres, testing of other boundary detection of algorithms for segmentation, testing rhythm features, testing more complex models instead of only a linear model, and not using the proposed approximation for factor scores.

Apart from exploring different methods, the significant results of this study also push for a follow-up to compare the proposed music thumbnailing method with other methods and Muziekweb's current implementation. The method proposed to evaluate the thumbnails is via another user study. In this user study, thumbnails obtained via different audio thumbnailing methods can be presented to participants. The respondent is then asked to rate these on their ability of being good thumbnails. This could be done similarly to this study via ranking. By comparing these thumbnails, it can be evaluated whether this method creates better thumbnails with regards to the current implementation used by Muziekweb, as well as how it holds up against other methods.

The results of this study also indicate that features for representativeness and catchiness do collude. The significance shows that it would interesting for further research to delve deeper into these two terms.

Chapter 6

Conclusion

Music thumbnailing is the process of selecting a fragment of a song that represents the complete song the best. Previous studies have mostly focused on detecting the chorus or the most repeated part for thumbnailing. Here a cognitive view is added by adding the notion of hooks, the most salient part of a musical piece. For this study, it is assumed that the hook overlaps with the most representative part. First, rankings of different fragments of the same song were obtained via a user study to compute the representative worth per fragment to grade. With features obtained via the CATCHY toolbox, the worth of each fragment was approximated. To reduce the dimensionality of the feature space and increase interpretability and identifiability, a five-factor model was made based on these CATCHY features via an Exploratory Factor Analysis. This resulted in intensity, commonality, and recurrence having a significant impact on the worth of a fragment. This colludes with musical aspects that have previously been identified as being important for the catchiness of the music. Simultaneously, the impact of familiarity on scoring the representativeness of fragments was tested and showed that familiarity is insignificant. Following, the influence of the segmentation method was also checked. The results indicate that the boundary detection algorithm used in this study does not perform too well. According to these findings, any arbitrary segmentation method can most likely be used to obtain fragments. Lastly, a new method to create music thumbnails is proposed based on the findings in this study.

References

Bartsch, M. A. and Wakefield, G. H. (2005). Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104. https://doi.org/10.1109/tmm.2004.840597.

Beggs, S., Cardell, S., and Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, 17(1):1–19. https://doi.org/10.1016/0304-4076(81)90056-7.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Burgoyne, J. A., Bountouridis, D., Van Balen, J., and Honing, H. (2013). Hooked: a game for discovering what makes music catchy. In *Proceedings of the 14th International Society of Music Information Retrieval Conference (ISMIR)*, pages 245–250, Curitiba, Brazil.

Burgoyne, J. A., Fujinaga, I., and Downie, J. S. (2016). Music information retrieval. *A New Companion to Digital Humanities*, pages 213–228.

Burgoyne, J. A., Wild, J., and Fujinaga, I. (2011). An expert ground truth set for audio chord recognition and music analysis. In *Proceedings of the 12th International Society of Music Information Retrieval Conference (ISMIR)*, volume 11, pages 633–638, Miami, USA.

Burns, G. (1987). A typology of 'hooks' in popular records. *Popular music*, 6(1):1–20. https://doi.org/10.1017/s0261143000006577.

Cano, P., Batlle, E., Kalker, T., and Haitsma, J. (2005). A review of audio fingerprinting. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, 41(3):271–284.

Chai, W. and Vercoe, B. (2003). Music thumbnailing via structural analysis. In *Proceedings of the 11th Association for Computing Machinery (ACM) International*

Conference on Multimedia, pages 223–226. https://doi.org/10.1145/957013.957057.

Conklin, D. and Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73. https://doi.org/10.1080/09298219508570672.

Cooper, M. L. and Foote, J. (2002). Automatic music summarization via similarity analysis. In *Proceedings of the 3rd International Society of Music Information Retrieval Conference (ISMIR)*, Paris, France.

Futrelle, J. and Downie, J. S. (2002). Interdisciplinary communities and research issues in music information retrieval. In *Proceedings of the 3rd International Society of Music Information Retrieval Conference (ISMIR)*, volume 2, pages 215–221, Paris, France.

Honing, H. J. (2010). Lure(d) into listening: The potential of cognition-based music information retrieval. 5(4):146–151. https://doi.org/10.18061/1811/48549.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., and Hothorn, M. T. (2016). Package 'multcomp'. *Simultaneous Inference in General Parametric Models. Project for Statistical Computing, Vienna, Austria*.

Huang, Y.-S., Chou, S.-Y., and Yang, Y.-H. (2017). Music thumbnailing via neural attention modeling of music emotion. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*, pages 347–350. IEEE. https://doi.org/10.1109/apsipa.2017.8282049.

Huron, D. B. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. MIT press. https://doi.org/10.7551/mitpress/6575.001.0001.

Jakubowski, K., Finkel, S., Stewart, L., and Müllensiefen, D. (2017). Dissecting an earworm: Melodic features and song popularity predict involuntary musical imagery. *Psychology of Aesthetics, Creativity, and the Arts*, 11(2):122–135. https://doi.org/10.1037/aca0000090.

Koops, H. V., de Haas, W. B., Bransen, J., and Volk, A. (2017). Chord label personalization through deep learning of integrated harmonic interval-based representations. In *Proceedings of the First International Workshop on Deep Learning and Music joint with International Joint Conference on Neural Networks (IJCNN)*, pages 19–25, Anchorage, US. arXiv preprint arXiv:1706.09552.

Kronengold, C. (2005). Accidents, hooks and theory. *Popular Music*, 24(3):381–397.

Levy, M. and Sandler, M. (2006). Application of segmentation and thumbnailing to music browsing and searching. In *Audio Engineering Society Convention 120*. Audio Engineering Society.

Müllensiefen, D. (2009). Fantastic: Feature analysis technology accessing statistics (in a corpus): Technical report v1. *London, England: Goldsmiths, University of London. Retrieved from http://www. doc. gold. ac. uk/isms/m4s/ Google Scholar*.

Müllensiefen, D. and Halpern, A. R. (2014). The role of features and context in recognition of novel melodies. *Music Perception: An Interdisciplinary Journal*, 31(5):418–435. https://doi.org/10.1525/mp.2014.31.5.418.

Müller, M., Jiang, N., and Grosche, P. (2013). A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(3):531–543. https://doi.org/10.1109/tasl.2012.2227732.

Nawata, H., Kamado, N., Saruwatari, H., and Shikano, K. (2011). Automatic musical thumbnailing based on audio object localization and its evaluation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp.2011.5946323.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear odels. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Nieto, O. and Bello, J. (2015). Lbd.30 msaf: Music structure analysis framework. In *Proceedings of the 16th International Society on Music Information Retrieval (ISMIR)*.

Osborne, J. W., Costello, A. B., and Kellow, J. T. (2008). Best practices in exploratory factor analysis. *Best Practices in Quantitative Methods*, pages 86–99. https://doi.org/10.4135/9781412995627.d8.

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Departent of Computing, City University, Londen, UK.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202. https://doi.org/10.2307/2346567.

Revelle, W. (2011). An overview of the psych package. *Department of Psychology Northwestern University*, 3:2012.

Schuller, B., Dibiasi, F., Eyben, F., and Rigoll, G. (2008). Music thumbnailing incorporating harmony- and rhythm structure. In *International Workshop on Adaptive Multimedia Retrieval*, pages 78–88. Springer.

Serra, J., Müller, M., Grosche, P., and Arcos, J. L. (2014). Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia*, 16(5):1229–1240. https://doi.org/10.1109/tmm.2014.2310701.

Silva, D. F., Falcao, F. V., and Andrade, N. (2018). Summarizing and comparing music data and its application on cover song identification. In *Proceedings of the 19th International Society of Music Information Retrieval Conference (ISMIR)*, Paris, France.

Tinsley, H. E. and Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, 34(4):414–424. https://doi.org/10.1037/0022-0167.34.4.414.

Turner, H. L., van Etten, J., Firth, D., and Kosmidis, I. (2018). Modelling rankings in r: the plackettluce package. *arXiv preprint arXiv:1810.12068*.

Van Balen, J. (2016). *Audio Description and Corpus Analysis of Popular Music*. PhD thesis, Utrecht University.

Van Balen, J., Burgoyne, J. A., Bountouridis, D., Müllensiefen, D., and Veltkamp, R. C. (2015a). Corpus analysis tools for computational hook discovery. In *Proceedings of the 16th International Society on Music Information Retrieval (ISMIR)*, pages 227–233, Malaga, Spain.

Van Balen, J., Wiering, F., and Veltkamp, R. (2015b). Audio bigrams as a unifying model of pitch-based song description. In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, Plymouth, United Kingdom.

Weihs, C., Jannach, D., Vatolkin, I., and Rudolph, G. (2016). *Music Data Analysis: Foundations and Applications*. CRC Press.

Williamson, V. J., Jilka, S. R., Fry, J., Finkel, S., Müllensiefen, D., and Stewart, L. (2012). How do "earworms" start? classifying the everyday circumstances of involuntary musical imagery. *Psychology of Music*, 40(3):259–284. https://doi.org/10.1177/0305735611418553.

Acronyms

CQT Constant-Q Transform. 25

EFA Exploratory Factor Analysis. 35, 36, 40, 44, 45, 51, 59, 61, 82–86, 88, 89

GLM Generalised Linear Model. 38, 39, 44, 46–51, 53–55, 60

HI Harmonisation Interval. 16, 33, 34, 52, 53, 86, 89

HIC Harmonic Interval Co-occurence. 16, 33, 34, 52, 86, 89

IClust Item Clustering Analysis. 36, 43, 80

INMI INvoluntary Musical Imargery. 12–14, 22, 56, 59

MFCC Mel-Frequency Cepstral Coefficient. 25, 33, 35, 53, 88, 89

MIB Melodic Interval Bigram. 16, 33, 34, 52, 53, 86, 89

MIR Music Information Retrieval. 8, 10, 12, 33

MSA Music Structure Analysis. 24, 25

MSAF Music Structure Analysis Framework. 24–26, 31, 35, 47, 48, 55, 57

PCA Principal Component Analysis. 11, 35, 51, 59, 86–89

PCP Pitch Class Profiles. 25

VSS Very Simple Structure. 36

Appendix A

Song Selection

Table A.1 displays the final selection of 60 songs used in this study. The artist, title, and the ID given by Muziekweb are shown. The Muziekweb ID is also used in further processing of the songs. The ID can be interpreted as stating the code for the album before the hyphen and the song identification after. As a great part of the list is derived from compilation albums, a great overlap of album ID's can be seen while the artists and their original albums do not overlap. These albums are JK155735, JK135666, JK135667, and JK135668 which are Dutch compilation albums with the 100 greatest hits in the 60's, 70's, 80's, and 90's respectively ("De 100 grootste jaren 60 hits", "De 100 grootste jaren 70 hits", "De 100 grootste jaren 80 hits", and "De 100 grootste jaren 90 hits"). The last compilation album with several songs in the list is HAX1853 contains a Dutch 100 best songs ("De 100 mooiste Nederlandstalige liedjes"). Behind the titles of the songs the language codes are added for Dutch (nl), English (en), French (fr), and German (de).

Table A.1: Selected tracks

Muziekweb ID	Artist	Title
HAX1853-0001	Marco Borsato	Wat is mijn hart (nl)
HAX1853-0002	De Dijk	Niemand in de stad (nl)
HAX1853-0004	Acda & De Munnik	Lopen tot de zon komt (nl)
HAX1853-0021	Guus Meeuwis	Toen ik je zag (nl)
HBX2087-0001	Gérard Lenorman	La ballade des gens heureux (fr)
HCX1490-0001	Peter Maffay	Du (de)
HCX1490-0005	Falco	Rock me Amadeus (de)
HEX14751-0004	Andrea Bocelli, Dua Lipa	If only (en)
JAX3145-0018	Memphis Minnie	Killer diller blues (en)
JK135666-0001	Meat Loaf	Paradise by the dashboard light (en)
JK135666-0002	Santana	She's not here (en)
JK135666-0003	Dr. Hook, Medicine Show	Sylvia's mother (en)
JK135666-0004	Cheap Trick	I want you to want me (en)
JK135666-0005	Redbone	We were all wounded at wounded Knee (en)
JK135666-0006	Boston	More than a feeling (en)
JK135666-0009	The Emotions	Best of my love (en)
JK135667-0001	Cyndi Lauper	Girls just wanna have fun (en)
JK135667-0002	Terence Trent D'Arby	Wishing well (en)
JK135667-0003	Bonnie Tyler	Total eclipse of the heart (en)
JK135667-0004	The Electric Light Orchestra	All over the world (en)
JK135667-0005	Alison Moyet	The ole devil called love (en)
JK135667-0006	Spandau Ballet	Through the barricades (en)
JK135667-0007	Earth, Wind & Fire	Let's groove (en)
JK135667-0009	Deniece Williams	Let's hear it for the boy (en)
JK135667-0010	Owen Paul	My favourite waste of time (en)
JK135667-0011	Bill Withers	Lovely day (en)
JK135667-0012	Spagna	Call me (en)
JK135667-0013	Fox the Fox	Precious little diamond (en)
JK135667-0014	Rodney Franklin	The groove (en)
JK135667-0015	Fiction Factory	Feels like heaven (en)

Continued on next page

Table A.1 { continued from previous page

Muziekweb ID	Artist	Title
JK135667-0017	Don Johnson	Heartbeat (en)
JK135667-0018	Shakin' Stevens	Oh Julie (en)
JK135667-0019	Toto	Africa (en)
JK135667-0020	Nicole	Don't you want my love (en)
JK135667-0021	Kenny Loggins	This is it
JK135667-0022	Journey	Don't stop believing
JK135667-0023	The Bangles	Eternal flame
JK135667-0024	The Outfield	Your love
JK135667-0025	The Weather Girls	It's raining men (en)
JK135667-0029	The Pasadenas	Tribute (en)
JK135668-0001	Des'ree	Life (en)
JK135668-0002	London Beat	I've been thinking about you (en)
JK135668-0003	Ten Sharp	You (en)
JK135668-0004	De Poema's	Mijn houten hart (nl)
JK135668-0005	Fugees	Fu-gee-la (en)
JK135668-0006	Michael Bolton	How am I supposed to live without you (en)
JK135668-0009	Martika (also in JK135667)	Toy soldiers (en)
JK135668-0011	Brownstone	If you love me (en)
JK135668-0013	B*Witched	C'est la vie (en)
JK135668-0018	Womack & Womack	Uptown (en)
JK135670-0002	The Pointer Sisters	Fire (en)
JK154457-0003	Big Blind	Hold on (en)
JK155735-0001	Bob Dylan	Like a rolling stone (en)
JK155735-0002	Simon & Garfunkel	Mrs. Robinson (en)
JK176222-0001	Actress	R.I.P. (en)
JK193539-0003	Ronnie Flex, Mr. Polska	Zusje (nl)
JK215477-0002	Boef, Sven Alias	Slapend Rijk (nl)
JK216338-0003	Josylvia, 3Robi, Killer Kamal	Westside (nl)
JK218358-0001	Lil' Kleine	Volume (nl)
JK234026-0003	Rosanne Cash, Elvis Costello, Kris Kristofferson	8 Gods of Harlem (en)

Appendix B

Informed Consent

Figure B.1 displays the informed consent for the user study as displayed in Qualtrics. The text translates roughly to:

> If you want to participate in this research, I ask of you to read the informed consent below thoroughly and to agree that you give your consent.
>
> Informed consent: I am informed to satisfaction about the research. I was given the possibility to ask questions about the research and possible questions have been answered to satisfaction. I have had the possibility to think about participation in the research. I understand that I am free to quit the experiment at any given moment. I understand that there are no risks or inconveniences to be expected due to participation in this experiment. I understand that the anonymous data that is obtained in this experiment will be digitally stored. I understand the obtained data will be used for scientific purposes and could be published. Hereby, I grant permission out of free will to participate in this experiment. Lastly, I confirm that I am at least 18 years old.

After the text, the page enforces the participants to choose between the option to give informed consent ("Ik geef toestemming") or deny ("Ik geef geen toestemming"). If the participant does not choose one of the options, a warning is given that denies them to proceed. If the participant give their informed consent, they are redirected to the rest of the survey, otherwise they are sent to the Qualtrics endscreen.

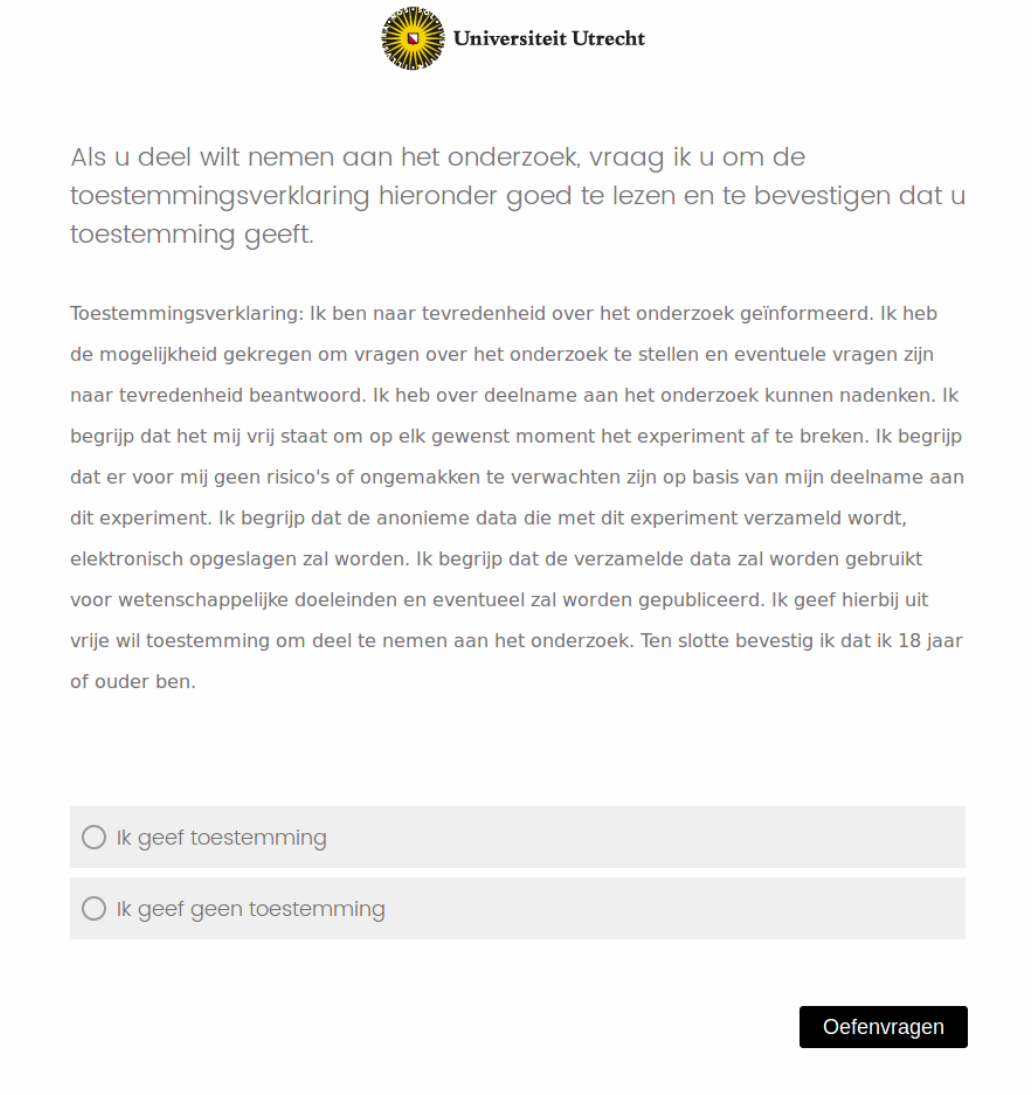

Figure B.1: Informed consent ("toestemmingsverklaring") as displayed in Qualtrics

Appendix C

Qualtrics

Figures C.1, C.2, C.3, C.4 and C.5 show screenshots of the Qualtrics survey as seen by participants. For each screenshot, a short explanation of which part of the survey is displayed as well as a translation to English of the Dutch text is given.

The survey starts with an introductory page as shown in Figure C.1 which states the general background of the study. The text can be translated to:

> Thank you for your interest! This study is a part of my master thesis "Artificial Intelligence" at the Utrecht University and is made possible by Muziekweb. I try to determine which fragment of a song conveys an idea of the complete song the best. I'd like to ask you to do the following: for several songs you will get to listen to three fragments. The title and artist are given for each song. You rank the fragments on how well they convey an idea of the song in your opinion. You do not have to be familiar with the songs. More information will be given in the next two examples. For any further questions you can mail to Arianne van Nieuwenhuijsen (a.n.vannieuwenhuijsen@uu.nl).

After clicking on the arrow to continue, the participant will be led to the informed consent which has already been discussed in Appendix B. After giving their consent, two pages with example question are shown with more details of the tasks. Figure C.2 shows the first of these example pages. The task as described here can be translated to:

> The next two songs are for practice.
>
> Rank the following three fragments on how well they convey an idea of the song. Drag the fragments in order of the best (1) to the least (3),

Figure C.1: Introduction to the survey as shown in Qualtrics

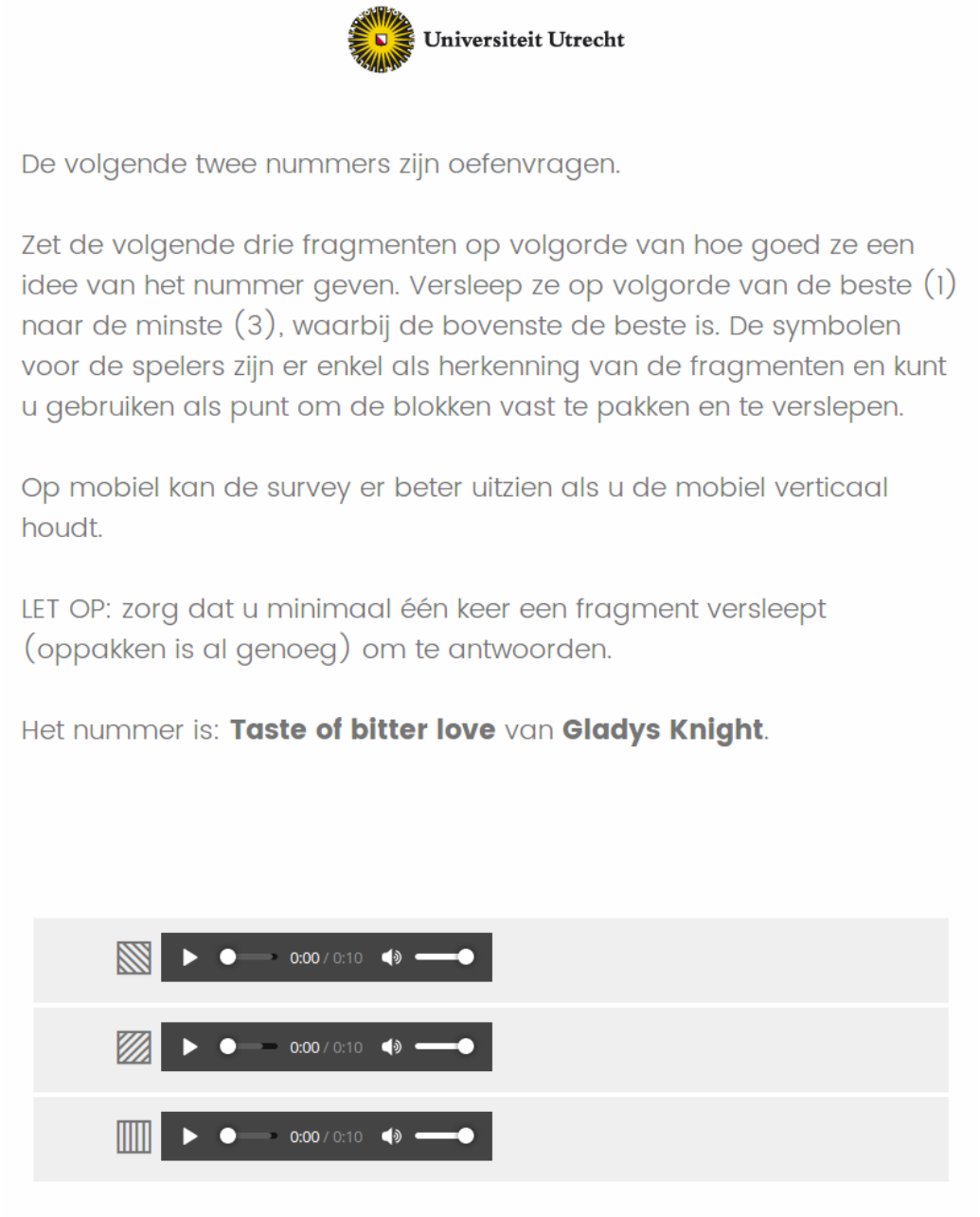

Figure C.2: An example question with instructions on the task as shown in Qualtrics

where the top is the best. The symbols in front of the players are only for recognition of the fragments and can be used as a mark to grab and drag the blocks.

On mobile, the survey could look better when you keep you phone vertically.

ATTENTION: ensure that you drag a fragment at least once (grabbing is enough) to save your answer.

The song is: Taste of bitter love by Gladys Knight

After the practice round, some last information regarding the survey is given as shown in Figure C.3):

These were the practice rounds. Now the real experiment starts. Each question takes about one minute. You may proceed as long as preferred and can stop at any given moment. If you restarts the survey within seven days on the same device, you can proceed the survey from where you left off. After seven days the survey is automatically closed. Thanks for your participation!

Figure C.3: The last page before the real experiment starts

Thereafter, the experiment starts with one of the randomly selected songs. After the practice rounds, it is assumed that the participant is familiar with the task, thus the text is kept minimal to keep distraction as low as possible as can be seen in Figure C.4. The example questions already mention that the survey might not look to well on mobile phone, as can be seen in Figure C.5. As the players in Qualtrics are not adjustable, it was chosen to leave the mobile version for what it is to bring out the survey as fast as possible and give a tip to hold your phone vertically. This might have influenced how many participants participated in the end.

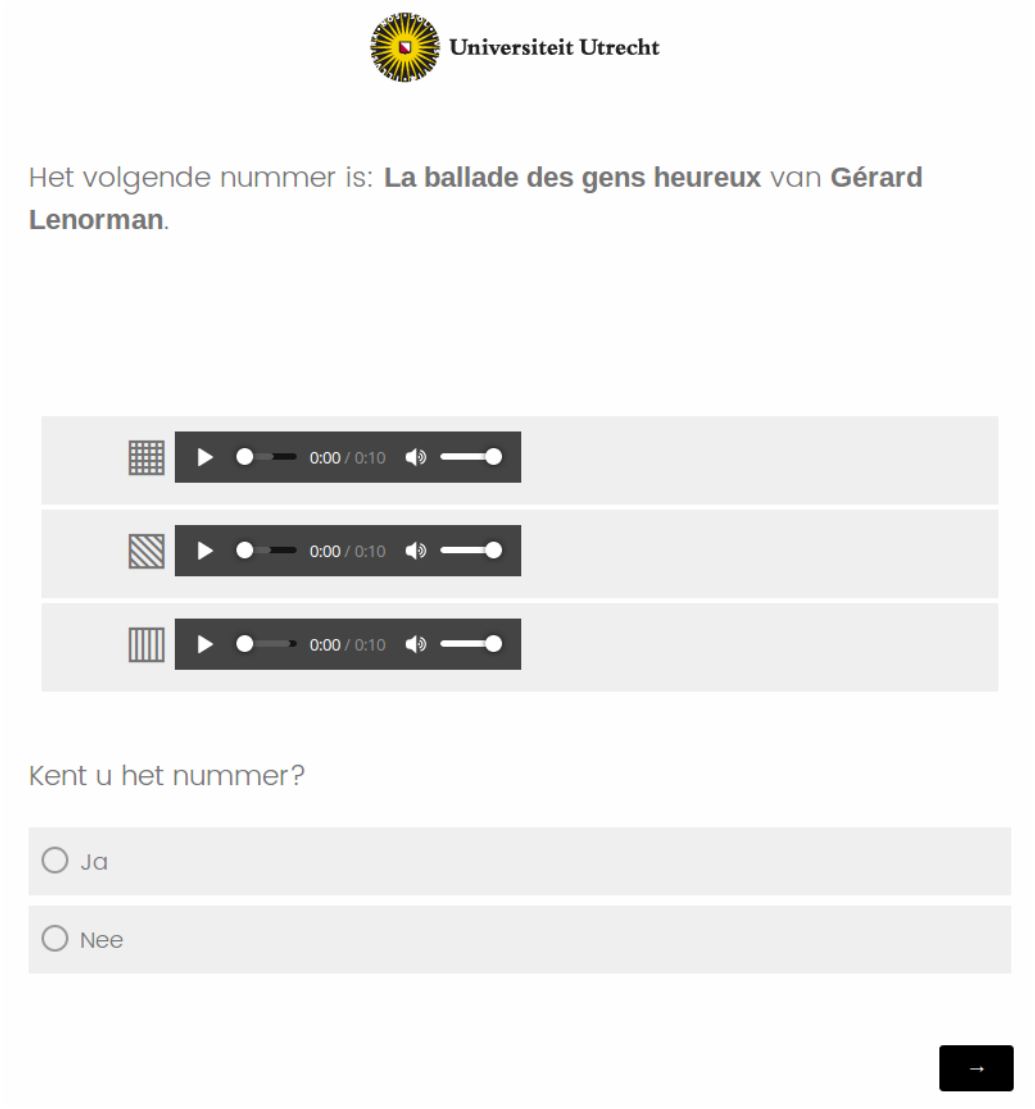

Figure C.4: One of the songs that is part of the main experiment as shown in Qualtrics

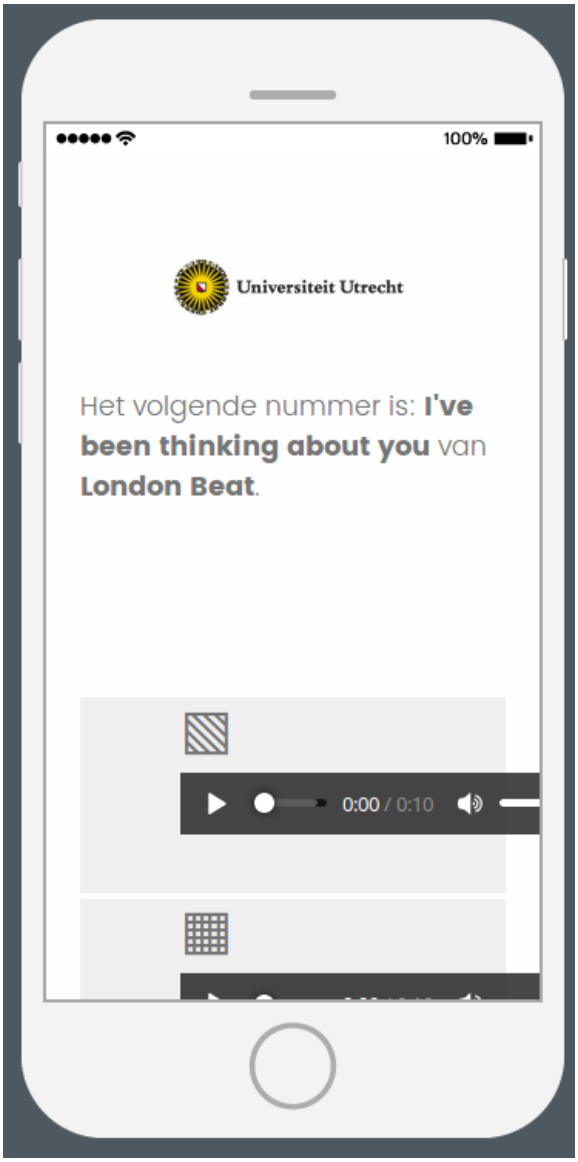

Figure C.5: Example of how the questions look on a mobile phone

Appendix D

IClust

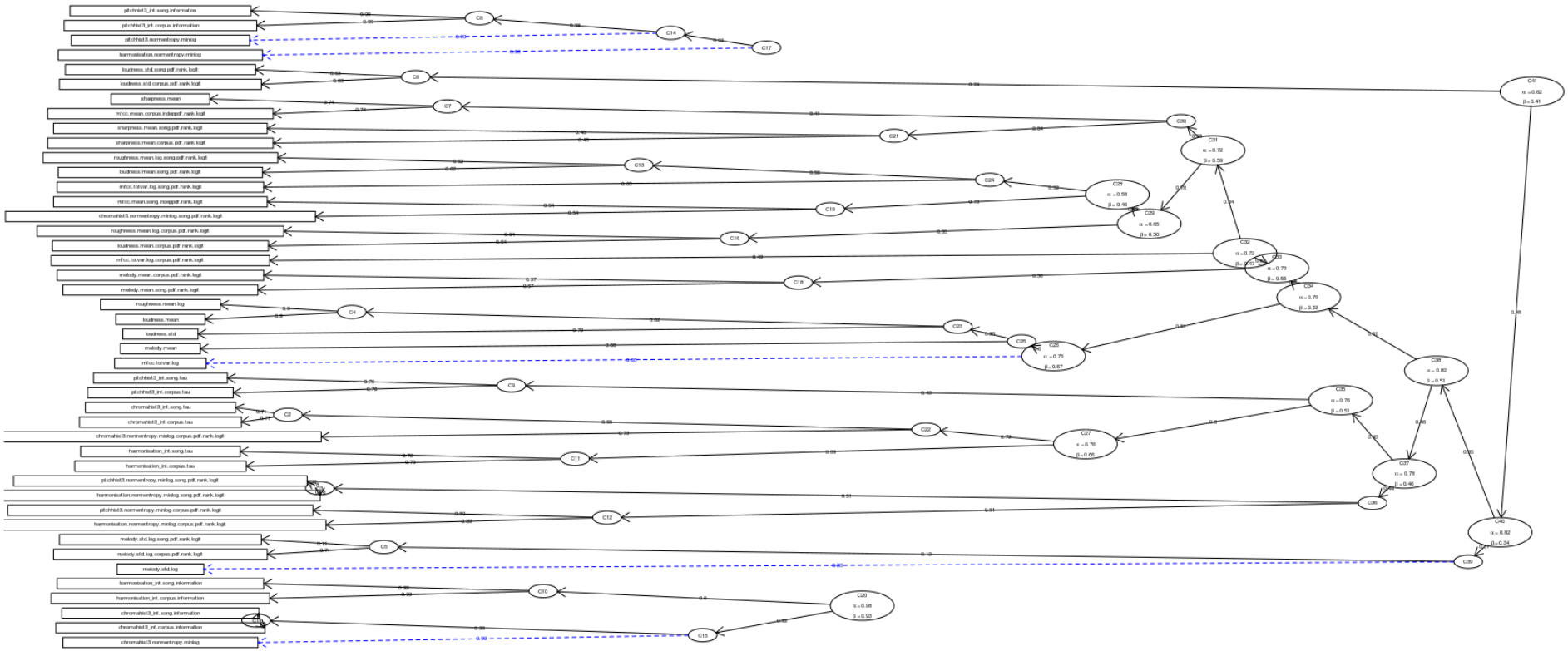

Figure D.1: Clustered identified by Item Clustering Analysis

Appendix E

Additional Factor Analysis Loadings

Tables E.1, E.2, E.3, and E.4 show the factor loadings for the minimum residual factor analysis with a number of 4, 6, 7, or 8 factors respectively.

	Factors						
Feature	MR1	MR2	MR3	MR4	h2	u2	com
HI *j* Song Information	**0.94**	0.06	-0.06	-0.05	0.89	0.11	1.02
HI *j* Corpus Information	**0.93**	0.07	-0.08	-0.06	0.88	0.12	1.03
HIC *j* Corpus Information	**0.92**	0.09	-0.07	0.01	0.86	0.14	1.03
HIC Entropy	**-0.91**	-0.05	0.09	-0.02	0.84	0.16	1.02
HIC *j* Song Information	**0.90**	0.09	-0.06	0.01	0.83	0.17	1.03
HI Entropy	**-0.78**	-0.03	**-0.42**	0.00	0.79	0.21	1.53
MIB *j* Corpus Information	**0.75**	-0.09	**0.51**	0.02	0.84	0.16	1.80
MIB Entropy	**-0.73**	0.06	**-0.57**	-0.03	0.87	0.13	1.92
MIB *j* Song Information	**0.73**	-0.10	**0.54**	0.02	0.83	0.17	1.89
HI *j* Corpus	**-0.51**	0.00	**0.31**	0.20	0.39	0.61	2.02
HIC Entropy *j* Corpus	**-0.44**	0.25	0.23	0.16	0.34	0.66	2.52
HI *j* Song	**-0.43**	0.06	**0.32**	0.23	0.35	0.65	2.50
HIC *j* Corpus	**-0.40**	0.19	**0.35**	0.21	0.36	0.64	3.02
HIC *j* Song	**-0.38**	0.26	0.29	0.17	0.33	0.67	3.15
Melodic Pitch SD	0.07	-0.05	-0.06	0.03	0.01	0.99	3.15
Loudness	0.02	**0.84**	-0.12	0.03	0.73	0.27	1.04
Roughness	0.19	**0.79**	0.03	0.09	0.67	0.33	1.15
Roughness *j* Corpus	0.07	**0.53**	0.27	0.03	0.35	0.65	1.54
MFCC Mean *j* Corpus	0.06	**0.51**	**0.34**	0.20	0.42	0.58	2.11
MFCC Variance	0.22	**-0.51**	-0.06	0.01	0.31	0.69	1.39
Loudness SD	0.30	**0.45**	-0.07	0.14	0.32	0.68	2.00
Melodic Pitch Height	0.13	**0.44**	-0.14	0.07	0.24	0.76	1.46
Roughness *j* Song	-0.01	**0.36**	0.30	0.04	0.22	0.78	1.96
Sharpness *j* Corpus	0.16	**0.36**	0.14	0.15	0.19	0.81	2.15
HIC Entropy *j* Song	-0.06	0.28	0.14	0.06	0.10	0.90	1.69
MFCC Mean *j* Song	0.00	0.26	0.24	0.16	0.15	0.85	2.65
Melodic Pitch Height *j* Corpus	0.07	0.26	0.17	0.14	0.12	0.88	2.60
Melodic Pitch Heigh *j* Corpus	-0.01	0.21	0.09	0.04	0.05	0.95	1.46
Sharpness *j* Song	0.08	0.17	0.15	0.06	0.06	0.94	2.71
MIB *j* Corpus	-0.09	0.07	**0.41**	0.30	0.27	0.73	2.00
Sharpness	**0.33**	0.28	**0.40**	0.17	0.38	0.62	3.16
Loudness *j* Corpus	0.09	0.02	**0.40**	-0.01	0.17	0.83	1.11
MIB *j* Song	0.04	0.06	**0.33**	**0.33**	0.22	0.78	2.10
Loudness *j* Song	0.00	0.10	**0.31**	0.07	0.11	0.89	1.32
Loudness SD *j* Corpus	0.04	0.11	0.30	0.02	0.10	0.90	1.30
MFCC Variance *j* Corpus	-0.02	0.05	0.26	-0.05	0.07	0.93	1.15
Melodic Pitch SD *j*Corpus	0.10	0.13	0.25	0.01	0.09	0.91	1.86
Loudness SD *j* Song	0.04	-0.01	0.21	0.00	0.05	0.95	1.08
MFCC Variance *j* Song	0.00	0.17	0.21	0.05	0.07	0.93	2.04
Melodic Pitch SD *j* Song	0.00	0.11	0.15	0.02	0.04	0.96	1.90
MIB Entropy *j* Corpus	-0.04	0.10	0.03	**0.78**	0.61	0.39	1.04
HI Entropy *j* Corpus	-0.05	0.08	0.07	**0.77**	0.61	0.39	1.05
HI Entropy *j* Song	-0.01	0.07	0.02	**0.51**	0.26	0.74	1.04
MIB Entropy *j* Song	-0.03	0.04	-0.05	**0.49**	0.24	0.76	1.05
SS loadings	7.82	3.49	3.07	2.28			

Table E.1: Loadings for a four-factor EFA model

Variable	MR1	MR6	MR3	MR4	MR2	MR5	h2	u2	com
HIC *j* Corpus Information	**0.94**	0.25	0.03	-0.01	0.05	0.00	0.95	0.05	1.15
HIC *j* Song Information	**0.93**	0.25	0.03	-0.02	0.08	-0.01	0.93	0.07	1.17
HIC Entropy	**-0.92**	-0.25	0.00	-0.01	0.00	-0.02	0.92	0.08	1.15
HI *j* Song Information	**0.85**	**0.37**	0.06	-0.04	-0.14	0.00	0.88	0.12	1.46
HI *j*Corpus Information	**0.85**	**0.35**	0.07	-0.03	-0.16	0.02	0.87	0.13	1.45
HI *j* Corpus	**-0.49**	-0.04	-0.03	0.13	**0.46**	-0.03	0.48	0.52	2.16
HIC Entropy *j* Corpus	**-0.43**	-0.07	0.22	0.12	0.28	0.12	0.34	0.66	2.76
HIC *j* Corpus	**-0.41**	0.03	0.14	0.15	**0.38**	0.12	0.37	0.63	2.73
HIC *j* Song	**-0.37**	-0.03	0.20	0.12	**0.35**	0.14	0.34	0.66	3.13
Melodic Pitch Height	0.10	-0.03	-0.06	0.01	0.04	-0.07	0.02	0.98	3.05
MIB Entropy	**-0.35**	**-0.93**	-0.01	-0.02	-0.06	-0.04	0.98	0.02	1.29
MIB *j* Song Information	**0.35**	**0.90**	-0.02	0.01	0.04	0.02	0.94	0.06	1.31
MIB *j*Corpus Information	**0.38**	**0.88**	-0.02	0.02	-0.01	0.05	0.93	0.07	1.37
HI Entropy	**-0.47**	**-0.79**	-0.09	0.00	0.00	-0.05	0.86	0.14	1.66
Melodic Pitch SD *j* Corpus	-0.02	0.28	0.17	-0.03	0.21	-0.07	0.16	0.84	2.76
Melodic Pitch SD *j* Song	-0.08	0.16	0.14	0.02	0.07	-0.01	0.06	0.94	2.84
Loudness	0.00	-0.02	**0.98**	0.05	-0.05	-0.03	0.97	0.03	1.01
Roughness	0.18	0.08	**0.78**	0.08	0.11	0.13	0.68	0.32	1.25
MFCC Variance	0.18	0.08	**-0.51**	0.04	-0.21	-0.01	0.35	0.65	1.70
Loudness SD	0.26	0.11	**0.48**	0.16	-0.09	0.06	0.34	0.66	2.05
Melodic Pitch Height	0.24	-0.13	**0.42**	0.03	0.17	-0.05	0.28	0.72	2.30
Sharpness *j* Corpus	0.15	0.10	0.30	0.13	0.17	0.17	0.19	0.81	3.62
MIB Entropy *j*Corpus	-0.08	0.05	0.12	**0.75**	0.12	-0.05	0.61	0.39	1.15
HI Entropy *j* Corpus	-0.08	0.05	0.09	**0.73**	0.17	-0.01	0.58	0.42	1.18
HI Entropy *j* Song	0.00	-0.03	0.02	**0.57**	-0.06	0.21	0.38	0.62	1.29
MIB Entropy *j* Song	-0.02	-0.06	0.02	**0.55**	-0.08	0.11	0.32	0.68	1.17
MIB *j* Corpus	-0.12	0.19	0.01	0.21	**0.54**	0.03	0.39	0.61	1.70
HI *j* Song	**-0.40**	-0.05	-0.01	0.15	**0.49**	0.06	0.43	0.57	2.20
MFCC Mean *j* Corpus	0.08	0.11	**0.39**	0.13	**0.44**	0.28	0.46	0.54	3.15
Sharpness	**0.31**	0.26	0.13	0.11	**0.38**	**0.36**	0.47	0.53	4.15
MIB *j* Song	-0.01	0.20	0.01	0.28	**0.35**	0.10	0.25	0.75	2.77
Loudness SD *j* Corpus	0.06	0.07	-0.02	-0.04	**0.31**	0.28	0.19	0.81	2.22
Loudness *j* Corpus	0.03	0.19	-0.13	-0.02	0.09	**0.55**	0.37	0.63	1.44
Roughness *j* Song	-0.06	0.10	0.26	0.05	0.04	**0.49**	0.33	0.67	1.72
Roughness *j* Corpus	0.09	0.04	**0.39**	0.01	0.20	**0.49**	0.44	0.56	2.37
Loudness *j* Song	-0.06	0.14	0.00	0.09	-0.01	**0.48**	0.26	0.74	1.27
MFCC Mean *j* Song	-0.05	0.11	0.19	0.17	0.07	**0.31**	0.18	0.82	2.81
MFCC Variance *j* Song	-0.05	0.10	0.11	0.07	0.01	0.29	0.12	0.88	1.75
HIC Entropy *j* Song	-0.07	0.02	0.22	0.07	0.03	0.25	0.13	0.87	2.38
MFCC Variance *j* Corpus	-0.11	0.17	0.01	-0.04	0.00	0.25	0.10	0.90	2.26
Melodic Pitch Height *j* Corpus	0.08	0.06	0.18	0.13	0.16	0.24	0.14	0.86	3.70
Loudness SD *j* Song	0.02	0.08	-0.08	-0.01	0.10	0.23	0.08	0.92	1.95
Melodic Pitch Height *j* Song	-0.04	0.04	0.18	0.07	-0.06	0.22	0.09	0.91	2.42
Sharpness *j* Song	0.03	0.13	0.14	0.06	0.04	0.16	0.07	0.93	3.52
SS loadings	5.96	4.01	3.12	2.11	2.03	1.99			

Table E.2: Loadings for a six-factor EFA model

Variable	MR1	MR6	MR3	MR2	MR4	MR7	MR5	h2	u2	com
HIC *j* Corpus Information	**0.95**	0.25	0.04	0.06	-0.02	0.05	-0.02	0.96	0.04	1.16
HIC *j* Song Information	**0.93**	0.25	0.03	0.07	-0.03	0.06	-0.05	0.94	0.06	1.18
HIC Entropy	**-0.93**	-0.25	0.00	-0.04	0.00	-0.01	-0.01	0.92	0.08	1.15
HI *j* Corpus Information	**0.85**	**0.36**	0.07	-0.03	-0.03	-0.14	0.04	0.87	0.13	1.44
HI *j* Song Information	**0.84**	**0.38**	0.06	-0.01	-0.02	-0.15	0.00	0.88	0.12	1.48
HI *j* Corpus	**-0.50**	-0.05	-0.03	0.19	0.11	**0.38**	-0.12	0.47	0.53	2.52
HIC Entropy *j* Corpus	**-0.47**	-0.05	0.19	**0.33**	0.16	0.08	-0.04	0.40	0.60	2.58
HIC *j* Corpus	**-0.44**	0.03	0.12	**0.34**	0.17	0.21	-0.04	0.40	0.60	2.96
HIC *j* Song	**-0.38**	-0.04	0.19	0.25	0.11	0.28	0.03	0.34	0.66	3.53
MIB Entropy	**-0.34**	**-0.92**	-0.01	-0.04	-0.01	-0.09	-0.04	0.98	0.02	1.30
MIB *j* Song Information	**0.34**	**0.91**	-0.03	0.07	0.03	0.00	-0.02	0.96	0.04	1.29
MIB *j* Corpus Information	**0.38**	**0.89**	-0.02	0.02	0.02	0.01	0.05	0.93	0.07	1.36
HI Entropy	**-0.46**	**-0.79**	-0.09	-0.04	-0.01	0.00	-0.03	0.86	0.14	1.65
Melodic Pitch SD *j* Corpus	-0.01	0.27	0.18	0.03	-0.05	0.22	-0.09	0.17	0.83	3.10
Loudness	0.01	-0.02	**1.00**	-0.02	0.04	-0.01	0.01	0.99	0.01	1.01
Roughness	0.17	0.08	**0.76**	0.20	0.09	0.04	0.07	0.67	0.33	1.33
MFCC Variance	0.18	0.09	**-0.51**	-0.13	0.05	-0.17	0.05	0.36	0.64	1.77
Loudness SD	0.25	0.12	**0.46**	0.09	0.19	-0.15	0.04	0.35	0.65	2.45
Melodic Pitch Height	0.23	-0.12	**0.40**	0.18	0.06	0.04	-0.15	0.29	0.71	2.70
Sharpness *j* Corpus	0.14	0.11	0.28	0.26	0.14	0.06	0.06	0.20	0.80	3.53
Sharpness	0.27	0.28	0.07	**0.60**	0.15	0.10	0.09	0.56	0.44	2.14
MFCC Mean *j* Corpus	0.04	0.13	**0.35**	**0.58**	0.17	0.14	0.00	0.53	0.47	2.12
Roughness *j* Corpus	0.07	0.05	**0.35**	**0.49**	0.02	0.01	0.28	0.45	0.55	2.58
Loudness SD *j* Corpus	0.03	0.08	-0.04	**0.40**	-0.03	0.12	0.08	0.19	0.81	1.42
Loudness *j* Corpus	0.01	0.20	-0.17	**0.40**	-0.02	-0.03	**0.39**	0.38	0.62	2.87
Melodic Pitch Height *j* Corpus	0.07	0.05	0.16	0.24	0.12	0.10	0.15	0.14	0.86	3.96
MFCC Variance *j* Corpus	-0.13	0.18	-0.01	0.20	-0.02	-0.09	0.16	0.12	0.88	4.23
HI Entropy *j* Corpus	-0.10	0.06	0.06	0.13	**0.79**	0.09	-0.04	0.67	0.33	1.15
MIB Entropy *j* Corpus	-0.09	0.05	0.11	0.06	**0.79**	0.09	-0.03	0.65	0.35	1.12
HI Entropy *j* Song	0.02	-0.05	0.03	-0.02	**0.52**	0.08	0.30	0.37	0.63	1.71
MIB Entropy *j* Song	0.00	-0.08	0.03	-0.10	**0.49**	0.08	0.24	0.32	0.68	1.67
MIB *j* Song	0.06	0.14	0.03	-0.03	0.18	**0.67**	0.18	0.54	0.46	1.42
MIB *j* Corpus	-0.11	0.15	0.01	0.19	0.16	**0.60**	-0.04	0.46	0.54	1.59
HI *j* Song	**-0.38**	-0.08	0.01	0.16	0.09	**0.54**	0.01	0.48	0.52	2.13
Melodic Pitch SD *j* Song	-0.06	0.15	0.16	-0.07	-0.02	0.18	0.05	0.09	0.91	3.76
Loudness *j* Song	-0.05	0.12	0.00	0.14	0.03	0.10	**0.49**	0.29	0.71	1.40
Roughness *j* Song	-0.05	0.09	0.26	0.22	0.00	0.09	**0.46**	0.35	0.65	2.34
MFCC Mean *j* Song	-0.03	0.08	0.21	0.09	0.11	0.20	**0.35**	0.23	0.77	2.87
MFCC Variance *j* Song	-0.03	0.08	0.12	0.06	0.01	0.12	**0.33**	0.15	0.85	1.83
Melodic Pitch Height *j* Song	-0.02	0.02	0.19	-0.01	0.02	0.07	0.28	0.12	0.88	1.92
HIC Entropy *j* Song	-0.07	0.02	0.22	0.14	0.05	0.04	0.22	0.12	0.88	3.08
Loudness SD *j* Song	0.02	0.08	-0.08	0.14	-0.04	0.09	0.19	0.08	0.92	3.43
Sharpness *j* Song	0.04	0.12	0.15	0.06	0.03	0.10	0.17	0.08	0.92	4.01
Melodic Pitch SD	0.09	-0.02	-0.06	0.02	0.03	-0.02	-0.10	0.02	0.98	3.17
SS loadings	5.93	4.03	3.02	2.09	2.06	1.76	1.44			

Table E.3: Loadings for a seven-factor EFA model

Variable	MR1	MR6	MR3	MR2	MR4	MR7	MR5	MR8	h2	u2	com
HIC *j* Corpus Information	**0.94**	0.26	0.04	0.05	-0.02	0.05	-0.02	-0.01	0.97	0.03	1.17
HIC *j* Song Information	**0.93**	0.26	0.04	0.06	-0.03	0.07	-0.05	-0.01	0.94	0.06	1.19
HIC Entropy	**-0.93**	-0.25	0.00	-0.04	0.00	0.00	0.00	0.00	0.92	0.08	1.15
HI *j* Corpus Information	**0.84**	**0.36**	0.07	-0.02	-0.03	-0.16	0.04	0.01	0.87	0.13	1.45
HI *j* Song Information	**0.84**	**0.38**	0.06	0.00	-0.02	-0.16	-0.01	0.01	0.88	0.12	1.50
HI *j* Corpus	**-0.50**	-0.04	-0.02	0.14	0.11	**0.43**	-0.14	0.00	0.48	0.52	2.43
HIC Entropy *j* Corpus	**-0.47**	-0.05	0.20	**0.31**	0.17	0.10	-0.08	0.01	0.40	0.60	2.71
HIC *j* Corpus	**-0.44**	0.04	0.14	0.30	0.17	0.25	-0.08	-0.01	0.40	0.60	3.17
HIC *j* Song	**-0.37**	-0.04	0.20	0.23	0.11	0.29	0.02	0.07	0.34	0.66	3.64
MIB Entropy	**-0.34**	**-0.93**	-0.01	-0.04	-0.01	-0.09	-0.04	-0.05	0.99	0.01	1.30
MIB *j* Song Indormation	**0.33**	**0.92**	-0.03	0.07	0.03	0.00	-0.03	0.03	0.96	0.04	1.28
MIB *j* Corpus Information	**0.37**	**0.89**	-0.02	0.03	0.02	0.00	0.05	0.04	0.94	0.06	1.36
HI Entropy	**-0.46**	**-0.79**	-0.09	-0.05	-0.01	0.01	-0.03	-0.06	0.85	0.15	1.67
Loudness	0.01	-0.03	**0.97**	-0.03	0.05	-0.03	0.03	0.08	0.96	0.04	1.03
Roughness	0.17	0.08	**0.78**	0.17	0.08	0.06	0.05	0.01	0.68	0.32	1.27
MFCC Variance	0.18	0.09	**-0.51**	-0.10	0.04	-0.18	0.05	-0.08	0.36	0.64	1.77
Loudness SD	0.24	0.13	**0.48**	0.07	0.18	-0.12	0.02	-0.08	0.36	0.64	2.29
Melodic Pitch Height	0.23	-0.13	**0.40**	0.15	0.07	0.04	-0.16	0.08	0.29	0.71	2.86
Sharpness *j* Corpus	0.14	0.10	0.29	0.25	0.14	0.06	0.03	0.02	0.20	0.80	3.48
HIC Entropy *j* Song	-0.07	0.02	0.22	0.16	0.05	0.04	0.20	-0.01	0.12	0.88	3.25
Sharpness	0.26	0.28	0.10	**0.59**	0.15	0.14	0.02	0.00	0.55	0.45	2.26
MFCC Mean *j* Corpus	0.03	0.14	**0.39**	**0.53**	0.16	0.21	-0.06	-0.06	0.53	0.47	2.69
Roughness *j* Corpus	0.07	0.04	**0.36**	**0.52**	0.04	0.01	0.23	0.05	0.46	0.54	2.29
Loudness *j* Corpus	0.01	0.18	-0.18	**0.49**	0.00	-0.08	**0.35**	0.10	0.45	0.55	2.63
Loudness SD *j* Corpus	0.04	0.06	-0.05	**0.43**	0.00	0.10	0.04	0.11	0.22	0.78	1.37
Melodic Pitch Height *j* Corpus	0.07	0.06	0.17	0.24	0.12	0.11	0.12	-0.02	0.14	0.86	3.86
MFCC Variance *j* Corpus	-0.13	0.18	0.00	0.21	-0.02	-0.08	0.13	-0.01	0.12	0.88	3.79
HI Entropy *j* Corpus	-0.10	0.06	0.06	0.11	**0.81**	0.08	-0.04	0.03	0.69	0.31	1.12
MIB Entropy *j* Corpus	-0.08	0.04	0.11	0.04	**0.80**	0.09	-0.02	0.02	0.67	0.33	1.10
HI Entropy *j* Song	0.02	-0.04	0.04	0.00	**0.50**	0.08	**0.31**	-0.04	0.36	0.64	1.79
MIB Entropy *j* Song	0.00	-0.07	0.04	-0.09	**0.47**	0.08	0.26	-0.05	0.32	0.68	1.81
MIB *j* Song	0.06	0.14	0.02	-0.03	0.18	**0.63**	0.22	0.11	0.51	0.49	1.64
MIB *j* Corpus	-0.10	0.15	0.01	0.16	0.16	**0.59**	-0.03	0.09	0.44	0.56	1.59
HI *j* Song	**-0.38**	-0.08	0.01	0.12	0.09	**0.57**	0.01	0.04	0.50	0.50	1.96
Loudness *j* Song	-0.05	0.12	0.01	0.19	0.02	0.10	**0.47**	-0.03	0.29	0.71	1.63
Roughness *j* Song	-0.06	0.10	0.27	0.26	-0.01	0.10	**0.43**	-0.01	0.35	0.65	2.76
MFCC Mean *j* Song	-0.03	0.10	0.22	0.09	0.09	0.22	**0.35**	-0.07	0.25	0.75	3.22
MFCC Variance *j* Song	-0.04	0.09	0.13	0.08	0.00	0.13	**0.32**	-0.03	0.15	0.85	2.09
Melodic Pitch Height *j* Song	-0.02	0.01	0.19	0.02	0.02	0.05	0.29	0.03	0.12	0.88	1.83
Loudness SD *j* Song	0.03	0.07	-0.09	0.17	-0.03	0.08	0.17	0.03	0.08	0.92	3.56
Sharpness *j* Song	0.04	0.12	0.15	0.06	0.03	0.09	0.16	0.02	0.08	0.92	4.11
Melodic Pitch SD	0.09	0.00	-0.05	0.00	0.02	0.02	-0.11	-0.09	0.03	0.97	3.47
Melodic Pitch SD *j* Corpus	0.01	0.20	0.12	0.06	0.01	0.09	-0.07	**0.76**	0.65	0.35	1.26
Melodic Pitch SD *j* Song	-0.04	0.07	0.11	-0.05	0.03	0.04	0.13	**0.60**	0.40	0.60	1.23
SS loadings	5.9	4.01	3.06	2.09	2.07	1.78	1.34	1.05			

Table E.4: Loadings for a eight-factor EFA model

Appendix F

Comparison of EFA with PCA

In this study, Exploratory Factor Analysis (EFA) is used as the preferred dimensionality reduction method. However, Van Balen et al. (2015a) used a Principal Component Analysis (PCA) for this purpose with CATCHY toolbox features. To see whether whether a different dimensionality reduction method produces a different division of features, the results of the PCA are compared to those of the EFA. PCA considers all variance whereas EFA only considers common, or shared, variance. This can result in a different interpretation of the variance within the data. Here, the results of a five-component PCA using Varimax rotation is shown to compare with the results of the EFA and Van Balen et al. As this comparison has no added value to the research question, this has not been included in the main body of text.

Table F.1 shows the loadings of the PCA model run for five components with Varimax rotation. To aid the comparison, the congruence between the different factors and components are shown in Table F.2 which gives an indication of how much factors and components overlap.

First, the components found as a result of the PCA in Table F.1 are interpreted by looking at the loadings above 0.5. This threshold differs from the other threshold in this paper, as it simplifies the interpretation by considering less features during the interpretation of the components. Simultaneously, the components are compared to the factors of EFA model and the twelve components found in the study Van Balen et al. (2015a). A first observation already shows that the ordering of features for the PCA already differs from the ordering for the EFA. This is because the ordering is based on the loadings per component or factor.

The first component consist of both the corpus and song-based second-order information features of HIC, HI, and MIB adding positively to the factor along with their first-order entropy variant negatively. This is similar to both MR1 and

	Components							
Feature	RC1	RC3	RC2	RC5	RC4	h2	u2	com
HI *j* Song Information	**0.90**	0.09	-0.24	-0.05	-0.03	0.88	0.12	1.17
HI *j* Corpus Information	**0.89**	0.10	-0.27	-0.03	-0.03	0.87	0.13	1.21
HIC *j* Corpus Information	**0.89**	0.11	-0.22	-0.05	0.02	0.85	0.15	1.16
HIC *j* Song Information	**0.88**	0.11	-0.19	-0.06	0.02	0.82	0.18	1.14
HIC Entropy	**-0.88**	-0.07	0.25	0.04	-0.03	0.84	0.16	1.18
HI Entropy	**-0.87**	-0.03	-0.21	-0.11	0.02	0.81	0.19	1.15
MIB *j* Corpus Information	**0.85**	-0.10	0.28	0.13	-0.02	0.83	0.17	1.29
MIB Entropy	**-0.84**	0.07	**-0.35**	-0.14	0.01	0.86	0.14	1.42
MIB *j* Song Information	**0.84**	-0.10	**0.33**	0.12	-0.02	0.83	0.17	1.38
Sharpness	**0.42**	0.23	0.26	**0.35**	0.14	0.43	0.57	3.59
HIC Entropy *j* Corpus	**-0.41**	0.25	**0.37**	0.16	0.08	0.40	0.60	3.08
Loudness	-0.01	**0.88**	-0.01	-0.02	0.03	0.77	0.23	1.00
Roughness	0.19	**0.80**	0.04	0.10	0.08	0.69	0.31	1.16
MFCC Variance	0.22	**-0.59**	-0.21	0.01	0.08	0.45	0.55	1.59
Melodic Pitch Height	0.11	**0.58**	-0.04	-0.17	0.05	0.38	0.62	1.29
Loudness SD	0.28	**0.54**	-0.13	0.00	0.19	0.42	0.58	1.94
MFCC Mean *j* Corpus	0.13	**0.51**	**0.33**	0.26	0.14	0.47	0.53	2.66
Rougness *j* Corpus	0.09	**0.49**	0.07	**0.48**	0.01	0.48	0.52	2.11
Sharpness *j* Corpus	0.19	**0.39**	0.10	0.15	0.15	0.24	0.76	2.31
MIB *j* Corpus	0.03	0.03	**0.62**	0.05	0.23	0.45	0.55	1.30
HI *j* Corpus	**-0.43**	-0.03	**0.56**	0.02	0.13	0.52	0.48	2.01
HI *j* Song	**-0.36**	0.01	**0.54**	0.10	0.17	0.46	0.54	2.08
HIC *j* Corpus	**-0.33**	0.16	**0.51**	0.16	0.14	0.44	0.56	2.35
Melodic Pitch SD *j* Corpus	0.20	0.18	**0.48**	-0.12	-0.15	0.34	0.66	2.01
HIC *j* Song	**-0.33**	0.24	**0.44**	0.16	0.11	0.40	0.60	2.95
MIB *j* Song	0.14	0.00	**0.43**	0.11	**0.34**	0.33	0.67	2.28
Melodic Pitch SD *j* Song	0.06	0.14	0.30	-0.05	-0.07	0.12	0.88	1.70
Loudness *j* Corpus	0.15	-0.14	0.10	**0.63**	-0.02	0.45	0.55	1.27
Loudness *j* Song	0.03	-0.04	0.02	**0.61**	0.11	0.38	0.62	1.08
Roughness *j* Song	0.00	0.28	0.04	**0.57**	0.05	0.41	0.59	1.49
MFCC Variance *j* Song	0.01	0.10	0.04	**0.40**	0.07	0.18	0.82	1.20
MFCC Mean *j* Song	0.02	0.20	0.09	**0.39**	0.21	0.25	0.75	2.26
MFCC Variance *j* Corpus	0.02	-0.03	0.12	**0.35**	-0.09	0.15	0.85	1.40
Loudness SD *j* Corpus	0.09	0.04	0.22	**0.35**	-0.07	0.19	0.81	1.97
HIC Entropy *j* Song	-0.06	0.26	0.01	**0.32**	0.07	0.18	0.82	2.12
Melodic Pitch Height *j* Song	-0.02	0.19	-0.07	0.30	0.08	0.14	0.86	2.00
Melodic Pitch Height *j* Corpus	0.10	0.25	0.07	0.26	0.18	0.18	0.82	3.22
Sharpness *j* Song	0.11	0.16	0.07	0.21	0.06	0.09	0.91	2.96
Melodic Pitch SD	0.07	-0.04	-0.03	-0.12	0.05	0.02	0.98	2.69
MIB Entropy *j* Corpus	-0.03	0.12	0.22	-0.07	**0.75**	0.62	0.38	1.24
HI Entropy *j* Corpus	-0.02	0.10	0.25	-0.04	**0.73**	0.61	0.39	1.28
HI Entropy *j* Song	-0.02	0.01	-0.08	0.17	**0.71**	0.54	0.46	1.14
MIB Entropy *j* Song	-0.05	0.00	-0.10	0.07	**0.70**	0.51	0.49	1.07
SS loadings	8.06	3.74	3.22	2.81	2.61			

Table F.1: Component loadings for PCA with five components

	MR1	MR2	MR3	MR4	MR5					
MR1	1.00	0.51	0.03	-0.15	-0.04					
MR2	0.51	1.00	0.15	0.14	0.32					
MR3	0.03	0.15	1.00	0.30	0.39					
MR4	-0.15	0.14	0.30	1.00	0.33					
MR5	-0.04	0.32	0.39	0.33	1.00	RC1	RC3	RC2	RC5	RC4
RC1	0.92	0.80	0.09	-0.03	0.12	1.00	0.10	-0.12	0.09	-0.03
RC3	0.07	0.09	0.99	0.29	0.39	0.10	1.00	0.19	0.26	0.22
RC2	-0.48	0.46	0.27	0.46	0.39	-0.12	0.19	1.00	0.33	0.27
RC5	-0.07	0.31	0.27	0.22	0.96	0.09	0.26	0.33	1.00	0.21
RC4	-0.09	0.05	0.22	0.97	0.30	-0.03	0.22	0.27	0.21	1.00

Table F.2: Congruence between the factors of the EFA and the components of the PCA

MR2 of the EFA and gives a first indication that the PCA prefers to divide features based on their derivation (whether via information, Kendall's , entropy, ranked density, or their first-order format) instead of what each feature measures in the audio signal. The very high loadings of above 0.8 also align very well with the features of component 1 in the study by Van Balen et al. (2015a).

The second listed component is RC3. It is apparent that while this component is the third in the model, it is more defined by very high loadings and thus listed second. The high loadings in this component are loudness and roughness which aligns with MR3 and could thus describe intensity. However, MFCC variance, melodic pitch height, the standard deviation of loudness, and the mean MFCC are also included in this component, which make it less clear to interpret. The first four high loadings do match those in component 2 in Van Balen et al.'s (2015a) study. Here the assumption that similarly derived features are clumped together is strengthened as these are the first-order variants of the basis features.

Then, RC2 seems to mix the type of derivation a little bit more, but describes mostly the Kendall's second-order features of the higher-dimensional features in the catchy toolbox. Interestingly, this component does not match one of the EFA factors and seems like a mix of components 6, 8, and 12 in the Van Balen et al. (2015a) study. The component seems to capture Kendall's and adds it positively, meaning that is describes commonness in the fragments according to the higher-dimensional melodic and harmonic features.

The fourth listed component is another switch based on the highness of the loadings. This component, RC5 is mostly composed of the second-order features of

loudness and roughness, with lower loadings for the second-order features of the two MFCC measures. This component seems to join factor MR5 in describing the conventionality of the intensity. In Van Balen et al. (2015a), component 7 seems to serve the same purpose.

Lastly, component RC4 contains the second-order variant of the MIB and HI entropy features. This agrees greatly with factor MR4 and components 4 and 10 in Van Balen et al. (2015a). Therefore, this component most likely depicts the conventionality of the melodic and harmonic features within the fragments.

The interpretation of these components indicate a great overlap between the factors of the EFA and the PCA. To strengthen this, Table F.2 shows the congruence between the factors and components. It can be seen that indeed, RC1, RC3, RC4, and RC5 have a high congruence with one of the EFA factors. This leaves RC2 which seems to be mostly based on how the second-order feature is computed and MR2 which does have a very high congruence with RC1 as well which is to be expected as RC1 contains HIC, HI, and MIB. Looking at the congruence between the factors and components themselves, it becomes apparent that the EFA results have far more overlap (highest congruence is 0.51), whereas the PCA results are more distinct with the highest congruence being 0.33. This could be due to difference in which variance is regarded. The choice to consider all or only shared variance could also have influenced the amount of very high loadings in the PCA in comparison to EFA. Considering all variance can cause inflated loadings (Osborne et al., 2008). Furthermore, the clumping together of second-order features with similar derivations in the PCA model seems less intuitive for interpretation, strengthening the choice for EFA as dimensionality reduction method.

In comparison with Van Balen et al. (2015a), it can be seen that the components here do match the components found in their study. The main difference is that using twelve components gives rise to more distinct components with a smaller amount of high loadings. In this study, the limit of five components causes features to cluster together far more. While a conscious choice has been made to ensure each factor in the EFA has at least three loadings of 0.4 or higher and the PCA was set up as similar as possible to the EFA, it might have not resulted in components that can be fully compared to those in Van Balen et al.. This means, for example, that in this study there is no component that is related to vocal prominence.