University Institute of Lisbon

Department of Information Science and Technology

Automated Generation of Movie Tributes

Ana Marta Simões Aparício

A Dissertation presented in partial fulfillment of the Requirements for the Degree of

Master in Computer Science

Supervisor

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Assistant Professor

ISCTE-IUL

Co-Supervisor

Doctor David Manuel Martins de Matos, Assistant Professor

Instituto Superior Técnico/Universidade de Lisboa

September 2015

"Somos do tamanho dos nossos sonhos."

Fernando Pessoa

Resumo

O objetivo desta tese é gerar um tributo a um filme sob a forma de *videoclip*, considerando como entrada um filme e um segmento musical coerente. Um tributo é considerado um vídeo que contém os clips mais significativos de um filme, reproduzidos sequencialmente, enquanto uma música toca. Nesta proposta, os clips a constar do tributo final são o resultado da sumarização das legendas do filme com um algoritmo de sumarização genérico. É importante que o artefacto seja coerente e fluido, pelo que há a necessidade de haver um equilíbrio entre a seleção de conteúdo importante e a seleção de conteúdo que esteja em harmonia com a música. Para tal, os clips são filtrados de forma a garantir que apenas aqueles que contêm a mesma emoção da música aparecem no vídeo final. Tal é feito através da extração de vetores de características áudio relacionadas com emoções das cenas às quais os clips pertencem e da música, e, de seguida, da sua comparação por meio do cálculo de uma medida de distância. Por fim, os clips filtrados preenchem a música cronologicamente. Os resultados foram positivos: em média, os tributos produzidos obtiveram 7 pontos, numa escala de 0 a 10, em critérios como seleção de conteúdo e coerência emocional, fruto de avaliação humana.

Palavras-chave: Artefato Multimédia, *Videoclip*, Seleção de Conteúdo, Sumarização de Texto, Coerência, Análise de Emoções.

Abstract

This thesis' purpose is to generate a movie tribute in the form of a videoclip for a given movie and music. A tribute is considered to be a video containing meaningful clips from the movie playing along with a cohesive music piece. In this work, we collect the clips by summarizing the movie subtitles with a generic summarization algorithm. It is important that the artifact is coherent and fluid, hence there is the need to balance between the selection of important content and the selection of content that is in harmony with the music. To achieve so, clips are filtered so as to ensure that only those that contain the same emotion as the music are chosen to appear in the final video. This is made by extracting vectors of emotion-related audio features from the scenes they belong to and from the music, and then comparing them with a distance measure. Finally, filtered clips fill the music length in a chronological order. Results were positive: on average, the produced tributes obtained scores of 7, on a scale from 0 to 10, on content selection, and emotional coherence criteria, from human evaluation.

Keywords: Multimedia Artifact, Videoclip, Content Selection, Text Summarization, Coherence, Emotion Analysis.

Acknowledgements

I would like to thank my family and friends for all the support. I would like to thank my colleges at L2F, specially Paulo Figueiredo, and Francisco Raposo, for being so helpful (and also precious friends). Thank you, Prof. Ricardo Ribeiro, and Prof. David Martins de Matos, for your loyal guidance.

Contents

List of Figures

Chapter 1

Introduction

This thesis focuses on the automatic generation of movie tributes. The intended output artifact consists of a short music video containing important parts of the movie playing along with the specified song. The length of the video corresponds to the song's. The purpose of a movie tribute consists of reliving emotions from the movie in a quick, and effective way. There are videos of this kind on YouTube, made manually by people who long to gather their most meaningful scenes so as to later remember a movie that they have seen and enjoyed very much.

The developed method has, as main concerns, content selection and coherence aspects.

It is essential to extract important content in order to generate a video that raises the viewer's interest, and to consider that the sequence of segments that composes a video artifact should correspond to parts that make-up a narrative following some intent (Branigan, 1992). If we want to collect the most important parts of the movie, we can consider the movie script, for instance, which offers us the narrative structure that the movie follows. Despite less informative, we can consider the subtitles as well. The summarization of these text sources can provide us a way to find the most important clips, through the timestamps of the output sentences a generic text summarization algorithm would provide.

Scripts typically include additional information in comparison with subtitles: apart from dialog, they include scenes descriptions and characters behavior. In fact, summarization of film scripts leads to better results than subtitles, however, their difference is not significant (Aparício et al., 2015).

Coherence is another important aspect to be considered when producing videos. If we want to ensure that we obtain a coherent video, the clips selected from the film should have a common ground. To achieve so, we can start by to considering using a centrality-based text summarization algorithm, that secures less diversity in its output. Furthermore, we can try filtering the selected clips to get only the ones that have specific emotions. For that purpose, we can consider the input music. In fact, music is known to have a profound effect on humans' emotions (Picard, 1997) and, for this reason, it is often used along with stories in order to emphasize their emotive content. So, if we compare the emotions transmitted by the music with the ones transmitted by each clip, we can choose only the clips which are more emotionally similar to the music to be presented in the final video. That can be performed extracting emotion-related features from the segments audio stream and the music, represent them as vectors, and compare them using a similarity measure.

1.1 Contributions

Two papers were produced and submitted to the arXiv, a repository of e-prints (electronic preprints) of scientific papers, which can be accessed online.

The first one is called "Summarization of Films and Documentaries Based on Subtitles and Scripts". Here we assess the performance of generic summarization algorithms when applied to subtitles and scripts, for films and documentaries. The performance of extractive summarization methods has been assessed in detail for news documents, for this reason we use the well–known behavior of news articles summarization as reference. We use three different datasets, the first composed of news, the second of fictional films subtitles and scripts, and the last one of non-fictional documentaries subtitles. The

evaluation is carried out with the standard ROUGE metrics, comparing human references, plot summaries, and synopses, against system-generated summaries.

Second, we produced a worked entitled "Generation of Multimedia Artifacts: An Extractive Summarization-based Approach", where we explore methods for content selection and address the issue of coherence in the context of the generation of multimedia artifacts. We use audio and video to present two case studies: generation of film tributes, and lecture-driven science talks. For content selection, we use centrality-based and diversity-based summarization, along with topic analysis. To establish coherence, we use the emotional content of music, for film tributes, and ensure topic similarity between lectures and documentaries, for science talks.

1.2 Demonstrations

Five tributes were generated, and presented in the 5th Lisbon Machine Learning School 2015 (on Demo Day), and in the "Noite Europeia dos Investigadores 2014 - 2015", on the National Museum of Natural History and Science, in Lisbon. The presented tribute were the following:

- "Atonement" (2007), with "La Plage", by Yann Tiersen;
- "300" (2006), with "To the Edge", by Lacuna Coil;
- "Furious Seven" (2015), with "See You Again", by Wiz Khalifa;
- "The Curious Case of Benjamin Button" (2008), with "The Last Goodbye", by Billy Boyd;
- "Interstellar" (2014), with "Conspiracy Agent", by Savant.

1.3 Thesis Structure

This thesis is organized as follows:

- Chapter 2 presents related work for multimedia artifacts, including content selection and coherence concerns;
- Chapter 3 presents an overview of generic summarization algorithms, describing LexRank and Support Sets on (centrality-based); Maximal Marginal Relevance (MMR) and Graph Random-walk with Absorbing StateS that HOPs among PEaks for Ranking (GRASSHOPPER) (diversity-based) and Latent Semantic Analysis (LSA);
- Chapter 4 presents an overview of Music Emotion Recognition (MER), including emotion description and music features used to identify emotions;
- Chapter 5 presents our proposed solution, including data processing, content selection, emotional coherence addressing and post-production concerns;
- Chapter 6 presents the dataset used in our experiments, our results and discussion;
- Chapter 7 presents our conclusions, contributions and directions for future research.

Chapter 2

Generation of Multimedia Artifacts

Automatic video generation has been explored in a wide variety of areas, including filmography (Brachmann et al., 2007), conference video proceedings (Amir et al., 2004), sports (Mendi et al., 2013), music (HUA et al., 2004), and matter-of-opinion documentaries (Bocconi et al., 2008). Specifically, the generation of multimedia artifacts has gained focus recently, as demonstrated by the emergence of techniques that try to create video summaries (Ding et al., 2012; Evangelopoulos et al., 2013; Ma et al., 2002a; Ding et al., 2012; Evangelopoulos et al., 2013; Irie et al., 2010), as well as techniques that use music as a means to produce videos (Nakano et al., 2011; HUA et al., 2004; Wu et al., 2012; Wang et al., 2012).

In this work, we need to assess content selection and coherence aspects, in order to create a movie tribute. In the following sections, we present some previous work concerning those matters.

2.1 Content Selection

Many approaches have used text to help determine important content (Ding et al., 2012; Evangelopoulos et al., 2013; Ma et al., 2002a). These are described in the following subsections.

2.1.1 Ding et al. (2012)

Ding et al. (2012) proposes the fusion of text, audio, and visual features, for multimedia summarization. This method includes extracting visual concept features and automatic speech recognition (ASR) transcription features from a given video, and developing a Template-Based Natural Language Generation System to produce a paragraph of natural language which summarizes the important information in a video belonging to a certain topic area, and provides explanations for why a video was matched, and retrieved based on the extracted features.

Visual concept features consist of 346 Motion Scale-Invariant Feature Transform (MOSIFT) (Chen and Hauptmann, 2009) and Coloured Scale-Invariant Feature Transform (CSIFT) features (Chen et al., 2013) that describe keyframes, originated by Support Vector Machine (SVM) classifiers trained over the Semantic Indexing (SIN) task in Text REtrieval Conference Video Retrieval Evaluation (TRECVID) 2011 Multimedia Event Detection (MED) (NIST, 2011). To determine the video-level semantic indexing, the average of the keyframe-level SIN is taken for all keyframes. This method includes the ranking of the detected visual concepts, in order to mention the most important ones in the recounting. It is then followed by the removal of less discriminative visual concepts, and re-ranking of the remaining visual concepts. Finally, the ground truth for each event is determined manually. Words spoken in a video are extracted using ASR and more weight is put on words which are semantically related to the description of the detected event, using the integration of WordNet (Miller, 1995) and Wikipedia (Kolb, 2009). Higher weights are assigned to unique words (words occurring more frequently in a particular event), determined based on the given positive samples in the development data. This system receives the features extracted from the video and triggers several Natural Language Generation (NLG) modules to generate text using pre-defined, static templates. The first module generates a general sentence about the topic the given video belongs to. The visual concept module generates several sentences about the objects and scenes observed in the video and re-ranks them. The top 5% visual concepts are picked, then compared with the topic signatures. The top visual concepts in the

event's signature are used in one to three recounting sentences; the last 50 visual concepts in the event's signature list are used in one to two recounting sentences. Some templates are generated to express the ranked ASR Transcription list in natural language. The "activity" module implements a grammar-based algorithm, which attempts to generate more relevant and complex sentences than the baseline visual concept module from frequently observed combinations of visual concepts.

2.1.2 Evangelopoulos et al. (2013)

Evangelopoulos et al. (2013) summarize movies using their subtitles (text), along with information from the audio and visual streams, integrating cues from these sources in a multi-modal saliency curve, using frame-level fusion (Equation 2.1, where m is the frame index).

$$S_{avt}[m] = \text{fusion}(S_a, S_v, S_t, m) \tag{2.1}$$

Auditory saliency (S_a) is determined by cues that compute multi-frequency waveform modulations; visual saliency (S_v) is calculated using intensity, color, and orientation values; and, textual saliency (S_t) is obtained through Part-Of-Speech (POS) tagging based on decision trees (Schmid, 1994), applied to the subtitles. A skimming percentage is predefined to create the final summary. First, it is created an attention curve using median filtering on the initial audio-visual-text (AVT) saliency. A saliency threshold T_c is selected so that only a c percentage of summarization is achieved: only frames m with saliency value $S_avt[m] > T_c$ are chosen to be included in the summary. This results in a video frame indicator function I_c that equals 1, $I_c[m]$, if frame m is selected for the summary, and 0, otherwise. I_c is then processed to form adjacent blocks of video segments, which involves eliminating small duration, isolated, segments and merging contiguous blocks of segments into one. Finally, the selected clips are joined using fade-in/fade-out for both the audio and visual streams.

2.1.3 Ma et al. (2002a)

Ma et al. (2002a) presents a method that models the user's attention in order to create video summaries. For a given video sequence, information is extracted from three channels: visual (object and camera motion, color, texture, shapes, and text regions), audio (speech, music, and various special sounds) and textual (obtained from closed caption, ASR, and superimposed text) to form three attention models. A linear combination is used to merge them (Equation 2.2, where w_v, w_a and w_l are the weights of the linear combination, M_v, M_a and M_t are normalized visual, audio, and textual attention models, respectively).

$$A = w_v \cdot \bar{M}_v + w_a \cdot \bar{M}_a + w_l \cdot \bar{M}_t \tag{2.2}$$

The resulting user attention curve is composed of a time series of the attention values associated with each frame in a video sequence. Based on the resulting curve, both keyframes and video skims are extracted, originating a straightforward shot-based approach to skim generation: once a skim ratio is given, skim segments are selected around each keyframe according to the skim ratio within a shot.

The number of needed keyframes in a shot and the dynamic skims are determined by the number of wave crests on attention curve, after smoothing and normalizing. A derivative curve is computed to discover crest peaks. The attention value of a keyframe is used as the importance measure of the keyframe, with which keyframes are ranked. For shot-based extraction, keyframes between two shot boundaries are used as representative frames of a shot. Its importance indicator is the maximum attention value of the keyframes in it. In case a shot presents no crest, the key-frame is considered the middle frame, and its importance value is zero. If only one keyframe is required for each shot, the keyframe with the maximum attention is selected; if the total number of keyframes allowed is less than the number of shots in a video, shots with lower importance values are ignored.

In order to avoid the interruption of speech within a sentence in an audio track, sentence boundary is done through an adaptive background sound level detection, involving the following steps: (1) Adaptive background sound level detection (used to set threshold for pause detection); (2) Pause and non-pause frame identification (using energy and zero-crossing-rate information); (3) Result smoothing based on the minimum pause length and the minimum speech length; (4) Sentence boundary detection (determined by longer pause duration).

2.2 Coherence

Coherence is an important aspect to be considering when producing videos. For instance, Irie et al. (2010) proposes a method that intends to generate a coherent movie trailer and tries to do so, re-ordering the set of the most emotional film segments. Initially, movies scenes are segmented into clips and each one is represented by a histogram of quantified audio-visual features related to emotions, such as color, image brightness, and motion intensity (Irie et al., 2009). Emotions are assigned to each clip considering its topics, retrieved with Latent Dirichlet Allocation (LDA) (Landauer et al., 1998) and a conditional probability table (CPT) containing emotional transition weights (Plutchik and Kellerman, 2013). Clips that represent emotions are extracted using a clustering method, named affinity propagation (AP) (Frey and Dueck, 2007), considering the Jensen-Shannon divergence (Melville et al., 2005) as the similarity function. Apart from that, it is considered that clips are similar if they are chronologically close. Each clip's duration is reduced proportionally to the trailer's. In order to optimize the reordering of the selected clips, it is used a method that estimates the affective impact (AI) of a clips sequence, based on a framework that calculates surprise, named Bayesian Surprise (BS), which determines the surprise induced in viewers when observing visual information (Itti and Baldi, 2006). Then, all shots are compared with each other, after calculating each shot emotional impact, so that emotionally similar shots are presented one after another.

2.2.1 Addressing Coherence Using Music

Music data has been explored as a mechanism to provide coherence in previous works (HUA et al., 2004; Wu et al., 2012). They are described in the following subsections.

2.2.1.1 HUA et al. (2004)

HUA et al. (2004) produces a short film to represent a song from personal home videos, based on repetitive visual and aural patterns of short films. Their system automatically extracts temporal structures of the video and music, as well as repetitive patterns in the music, and tries to match important segments from the raw home video footage accordingly.

First, the raw home video is segmented into shots, according to color similarities or time-stamp. Then, an attention measure (importance) is attributed to each shot, the result of the average attention index of each video frame (Ma et al., 2002b). This value is related to object motion, camera motion, color, and speech in the video. Type and speed of camera motion, motion intensity, and color entropy (information) are also retrieved using the results of the above analysis. Onset series are extracted to estimate the music's rhythm, align music clip and video shot boundaries, and discover the corresponding onset strengths in the incidental music. This is done by analyzing energy peaks in the frequency domain. Temporal, spectral and Constant Q Transform (CQT) features are extracted to identify repeating patterns and structure. Temporal features are used to estimate tempo, period, and the length of a musical phrase (used as the minimum length of a significant repetition in repeating patterns discovery and boundary determination). Spectral features are used for vocal and instrumental sound discrimination, as well as to identify the prelude, interlude, and coda of the song. CQT features are used to represent the note and melody information. Based on these features, a self-similarity matrix of the music is obtained, from which the significant repeating patterns are detected, with an adaptive threshold setting method (Equation 2.3, where $0 \leq i < K$, K is the number of patterns (prelude, interlude, and coda are regarded as one pattern called "instrument");

$Type_i$ and Num_i are the type (normal or instrument, the former one covers all non-instrument patterns); ($Start_{ij}$, End_{ij}, $Tempo_{ij}$) are the start time, end time and tempo of j-th occurrence of a music pattern MP_i in the music). The boundaries of repeating patterns are aligned using an optimization-based approach that uses the obtained structure.

$$MP_i = (Type_i, Num_i, MS_{ij} = (Start_{ij}, End_{ij}, Tempo_{ij}), 0 \leq j < Num_i) \quad (2.3)$$

To find an appropriate set of shots from the video, for each music repetitive pattern, shots with very low color entropy or extremely high camera motion speed (low-quality shots) are removed from the shot list, as done by Hua et al. (2004a). Scene segmentation is done by grouping the shots according to content similarity and time-stamp (if available). The similarity function is given by the weighted sum of a series of histogram intersections of both shots.

In order to select appropriate video segments for each music segment, as well as align shot transitions with the strong onsets in music, the corresponding music segment of every occurrence of the specific music pattern is divided into small music sub-clips by finding strong onsets in a sliding window. To guarantee visual consistency for a certain music pattern, the corresponding video segments are selected from the same scene, while those for different occurrences of the same music pattern are selected from different segments of the shots in the assigned scene, to promote visual content variety along the time-line. However, this may introduce inconsistency.

2.2.1.2 Wang et al. (2012)

Wang et al. (2012) presents a novel machine learning model to learn the tripartite relationship among music, video, and emotion simultaneously, from an emotion-annotated corpus of music videos, in order to bridge music and video. This model is applied to predict emotion distributions in a stochastic emotion space from low-level acoustic

features. The music's and video's emotions are met by comparing their emotion distributions, using a similarity measure, the Euclidean distance.

It is proposed a novel acoustic emotion Gaussian (AEG) model that learns two sets of Gaussian mixture models (GMM) from data, namely acoustic GMM and Valence/Arousal (VA) GMM. The acoustic GMM computes low-level acoustic features, such as loudness, timbre, rhythm, and harmony, and the VA GMM describes high-level emotions. In order to align the two GMMs, the system performs semantic mappings between the acoustic feature space and the music emotion space, introducing a set of latent feature classes, $z_k{}_{k=1}^{K}$. Each z_k is defined by a latent acoustic classifier A_k, that maps a specific pattern of acoustic features to a specific area G_k in the VA space. The set of latent acoustic classifiers, $A_k{}_{k=1}^{K}$, can be implemented by a universal acoustic GMM, in which A_k represents a specific acoustic pattern discovered by the GMM learning in the frame-based acoustic feature space. G_k can be modeled by a bivariate Gaussian distribution, becoming a latent VA Gaussian. The VA GMM corresponds to the mixture of latent VA Gaussians.

In the process of generation of music emotion, the acoustic features of a music clip are represented by computing the posterior probabilities over the acoustic GMM (each Gaussian component A_k leads to a posterior probability θ_k, based on its frame-based feature vectors). This clip-level acoustic feature representation is the acoustic GMM posterior, $\theta_{k=1}^{K}$, subject to $\sum_k \theta_k = 1$, which captures the acoustic characteristics of every music clip in a K-dimensional probabilistic space. The emotion distribution of a music clip in the emotion (VA) space can be generated by the weighted combination of all latent VA Gaussians as $\sum_k \theta_k G_k$ using $\theta_k{}_{k=1}^{K}$ as weights.

The emotion-based music retrieval is divided into two phases: the feature indexing phase and the music retrieval phase.

In the indexing phase, each music clip in the unlabeled music database is indexed with two indexing approaches based on the music clip's acoustic features: (a) using the acoustic GMM posterior (a fixed-dimensional vector) of a clip using the acoustic GMM, or (b) using the predicted emotion distribution (a single 2-D Gaussian) of a clip given by automatic emotion prediction.

In the retrieval phase, given a point query from the VA space, the system will return a ranked list of relevant music clips. Two matching methods are applied, namely (a) pseudo song-based matching (first indexing approach) and (b) distribution likelihood-based matching (second indexing approach).

In the first indexing approach, the point query is first transformed into a pseudo song, the estimated acoustic GMM posterior: it is transformed into probabilities ${\lambda_k}_{k=1}^{K}$, s.t. $\sum_k \lambda_k = 1$. The resulting λ_k represent the importance of the k-th latent VA Gaussian for the input query point. The pseudo-song is then matched with clips in an unlabeled music database.

In the second indexing approach, a point query is fed into the predicted emotion distribution of each clip in an unlabeled music database, and the system ranks all the clips according to the estimated likelihoods.

To start the generative process of the AEG model, it is used a universal acoustic GMM to span a probabilistic space with a set of diverse acoustic Gaussians. To cover the emotion perception of different subjects, typically, a clip is annotated by multiple subjects, using a user prior model to express the contribution of each individual subject. Finally, in response to the query, the retrieval system ranks all the clips in descending order of cosine similarity.

2.3 Summary

In this chapter, we explored previous work on multimedia artifacts generation, including methods developed to address issues such as content selection and coherence of the final video.

Chapter 3

Generic Summarization

In this chapter, we present five text-based summarization approaches, to explore the possibilities we have to summarize the movie's subtitles.

3.1 Generic Text Summarization Algorithms

Several generic summarization algorithms have been developed to determine relevant content. In this section, we present five text-based summarization approaches: LexRank (Erkan and Radev, 2004) and Support Sets (Ribeiro and de Matos, 2011), which are centrality-based, MMR (Carbonell and Goldstein, 1998) and GRASSHOPPER (Zhu et al., 2007), which are diversity-based, and LSA (Gong and Liu, 2001).

3.1.1 Centrality

Centrality-based algorithms consider that the most important content of an input is the most central, considering its representation as a graph, spatial, etc.

3.1.1.1 LexRank

LexRank (Erkan and Radev, 2004) is a centrality-based method based on Google's PageRank for ranking web pages (Brin and Page, 1998). A graph is built using sentences, represented by TF-IDF score vectors, as vertexes and cosine similarity is used to determine how they connect. An edge is created if the similarity score exceeds some threshold. Then, the calculation described by Equation 3.1 is computed for each vertex until convergence, which happens when the error rate between two successive iterations is lower than a certain value for each vertex.

$$S\left(V_i\right) = \frac{\left(1-d\right)}{N} + d \times \sum_{V_j \in \mathrm{adj}[V_i]} \frac{\mathrm{Sim}\left(V_i, V_j\right)}{\sum_{V_k \in \mathrm{adj}[V_j]} \mathrm{Sim}\left(V_j, V_k\right)} S\left(V_j\right) \tag{3.1}$$

where d is a damping factor which ensures the convergence of the method, N is the total number of vertexes, and $S\left(V_i\right)$ is the score of the ith vertex.

3.1.1.2 Support Sets

Documents typically are composed by a mixture of subjects, normally involving a main subject and other minor (lateral) issues. Support Sets are defined based on this idea (Ribeiro and de Matos, 2011). Important content can be determined by creating a support set for each passage of the input, determined by comparing each passage with all remaining ones from the source. The most semantically-related passages are included in the support set, determined via geometric proximity. In this manner, groups of related passages are uncovered, each one representing a topic. A summary can be composed by selecting the most relevant passages, which are the ones present in the largest number of support sets.

Given a segmented information source $I \triangleq p_1, p_2, \ldots, p_N$, support sets S_i associated with each passage p_i are defined as indicated in Equation 3.2, where Sim represents a similarity function, and ϵ_i is a threshold.

$$S_i \triangleq \{s \in I : \mathrm{Sim}(s, p_i) > \epsilon_i \wedge s \neq p_i\} \tag{3.2}$$

The most important passages are selected based on Equation 3.3.

$$\arg \max_{s \in U_{i=1}^{n} S_i} |\{S_i : s \in S_i\}| \tag{3.3}$$

3.1.2 Diversity

A very common problem in automatic summarization is the presence of redundant information in the final summary. In order to solve this issue, the following algorithms were developed to guarantee diversity.

3.1.2.1 Maximal Marginal Relevance (MMR)

MMR is a commonly adopted method for query-focused summarization (Carbonell and Goldstein, 1998). A linear combination of relevance and novelty is established by configuring the model, which iteratively selects the documents that result from applying Equation 3.4, where Sim_1 and Sim_2 represent similarity metrics. S_i represents the non-selected documents and S_j the previously selected documents, Q is the query and λ is the parameter that configures between relevance, $\lambda(\mathrm{Sim}_1(S_i, Q))$, and novelty, $(1-\lambda)(\max_{S_j} \mathrm{Sim}_2(S_i, S_j))$.

$$\arg \max_{S_i} \left[\lambda \left(\mathrm{Sim}_1 \left(S_i, Q \right) \right) - (1 - \lambda) \max_{S_j} \mathrm{Sim}_2 \left(S_i, S_j \right) \right] \tag{3.4}$$

Therefore, when $\lambda = 1$, the summary will be composed by the standard relevance list. On the other hand, for $\lambda = 0$, maximal diversity ranking is obtained. A good practice is to first observe the information space surrounding the query with $\lambda \simeq 0.3$, and then focus on the important parts applying MMR with $\lambda \simeq 0.7$ (Carbonell and Goldstein, 1998). Furthermore, the MMR approach can generate generic summaries by

considering the input sentences centroid as a query (Murray et al., 2005; Xie and Liu, 2008).

3.1.2.2 Graph Random-walk with Absorbing StateS that HOPs among PEaks for Ranking (GRASSHOPPER)

GRASSHOPPER (Zhu et al., 2007), is a graph-based ranking algorithm, which focuses on maximizing diversity while minimizing redundancy. GRASSHOPPER is based on random walks in an absorbing Markov chain and can be seen as an alternative to MMR, because it includes diversity as well. The algorithm receives three parameters as input: a weighted graph W, a probability distribution r (user-defined prior ranking) and $\lambda \in [0, 1]$ that balances the weighted graph and the prior ranking. W is a $n \times n$ matrix, where n represents sentences, and the weights can be defined using a similarity measure such as the cosine distance. r is a user-defined ranking (e.g., sentence position), defined by $r = (r_1, \ldots, r_n)$, where $\sum_{i=1}^{n} r_i = 1$, and r_i is the probability of sentence i. When there is no prior ranking, a uniform distribution can be used, where each sentence has the same probability.

Sentences are ranked by applying the teleporting random walks method, which is based on the $n \times n$ transition matrix $\tilde{P}$ (calculated by normalizing the rows of W):

$$P = \lambda \tilde{P} + (1 - \lambda)\, \mathbf{1}\mathbf{r}^{\top} \tag{3.5}$$

The first sentence to be scored is the one with the highest stationary probability $\arg\max_{i=1}^{n} \pi_i$ according to the stationary distribution of P: $\pi = P^{\top}\pi$.

The already selected sentences may never be visited again, by defining $P_{gg} = 1$ and $P_{gi} = 0, \forall i \neq g$. The expected number of visits is given by an N matrix, as defined by Equation 3.6, where N_{ij} is the expected number of visits to the sentence j, if the random walker began at sentence i.

$$N = (I - Q)^{-1} \tag{3.6}$$

We then obtain the average of all possible starting sentences to get the expected number of visits to the jth sentence, v_j. The sentence to be selected is the one given by the following equation:

$$\arg \max_{i=|G|+1}^{n} v_i \tag{3.7}$$

3.1.3 Latent Semantic Analysis (LSA)

LSA is a mathematical technique based on Singular Value Decomposition (SVD). This approach aims at inferring contextual usage of text based on word co-occurrence. Gong and Liu (2001) use LSA for determining important topics in documents without the need of external lexical resources, such as an online thesauri or dictionaries. This technique follows the notion that the occurrence context of a particular word provides information that can be used to assess meaning. LSA produces relations between words and sentences that correlate with the way humans make associations or discover semantic similarity (Landauer and Dutnais, 1997; Landauer et al., 1998).

In order to use LSA for text summarization, the input document is represented through a $t \times n$ term-by-sentences matrix A, where rows represent unique words (t) of the input document, and columns represent sentences (n). Then, SVD is applied to A, resulting in its decomposition (Equation 3.8): U, a $t \times n$ matrix of left singular vectors (its columns); Σ, an $n \times n$ diagonal matrix of singular values; V^T, an $n \times n$ matrix of right singular vectors (its rows).

$$A = U\Sigma V^T \tag{3.8}$$

Some of the singular values in the diagonal matrix Σ are too small and can be discarded, by setting them to zero. Therefore, by keeping the first, non-zero entries, of singular values of Σ, we reduce it to Σ_k, which represents the $k \times k$ submatrix of Σ. U and V^T are also reduced to U_k and V_k^T in order to have k columns and k rows,

respectively. Therefore, A is approximated to: $A_k = U_k \Sigma_k V_k^T$, an $m \times n$ matrix, U is $m \times k$, Σ is $k \times k$ and V is $k \times n$.

When SVD is applied to matrix A, the transformation can be seen from two perspectives: as a mapping from one dimensional space to another, resulting in dimensionality reduction; and a semantic structure derived by aggregating words and sentences in similar contexts. If a pattern of words occurs very often in a document, this pattern will be captured and represented by one of the singular vectors, which represent topics or concepts, following the notion that similar words appear in similar contexts.

3.2 Summary

In this chapter, we presented five summarization approaches, which have been successfully applied to text. These algorithms extract the most relevant and/or diverse information from the input, according to each algorithm's definition of relevance and diversity. This is done by scoring and ranking sentences and then picking the ones with highest scores to include in the summary.

Chapter 4

Music Emotion Recognition (MER)

One of the main issues addressed in this thesis is the coherence of our final multimedia artifact. We use emotions to filter the clips resulting from the content selection phase. Specifically, we extract emotion-based features from the clips' auditory channel to perceive the present emotions, and compare them with the emotion-based features of the music piece, and consider only the most similar to the latter to compose the final video. This chapter presents an overview of emotion characterization and emotion identification, including types of emotion-related features that have been explored to perceive emotions in music, some of which we used in this thesis.

4.1 Emotion Characterization

Languages of emotion include music as "one of the finest" (Picard, 1997).

Psychologists have been studying the relationship between music and emotion for the past decades, facing the important problem of conceptualization of music emotion (Juslin and Sloboda, 2011). Two approaches have emerged that try to conceptualize emotion: the first is a category-based; the second is a dimension-based approach. They emerged from empirical studies, where people's verbal reports of emotion responses are considered.

4.1.1 Categorical Approach

The categorical approach to emotion conceptualization is based on the concept of the existence of a limited set of basic, innate and universal emotions, such as happiness, sadness, anger, fear, disgust, and surprise. This approach groups emotions into categories from which all other emotion classes can be derived (Anderson and McOwan, 2006; Ekman, 1992; Picard et al., 2001; Schuller et al., 2010). Schubert's nine clusters and the correspondent emotional adjectives (Schubert, 2003) (Table 4.1) is an example of such approach.

TABLE 4.1: The Nine Emotion Clusters Proposed by Schubert (2003)

Cluster	Emotions in Each Cluster
1	Bright, cheerful, happy, joyous
2	Humorous, light, lyrical, merry, playful
3	Calm, delicate, graceful, quiet, relaxed, serene, soothing, tender, tranquil
4	Dreamy, sentimental
5	Dark, depressing, gloomy, melancholy, mournful, sad, solemn
6	Heavy, majestic, sacred, serious, spiritual, vigorous
7	Tragic, yearning
8	Agitated, angry, restless, tense
9	Dramatic, exciting, exhilarated, passionate, sensational, soaring, triumphant

4.1.2 Dimensional Approach

The dimensional approach to emotion conceptualization identifies different emotion "dimensions" that correspond to internal human representations of emotion, which are

found by analyzing the correlation between emotional terms. These dimensions are represented as axes in a 2D plan. It differs from the categorical approach, which focuses on characteristics that distinguish emotions between them.

Russell proposed a two-dimensional circular structure as a model of affect (Russell, 1980), that identifies valence and arousal as emotion dimensions (Figure 4.1).

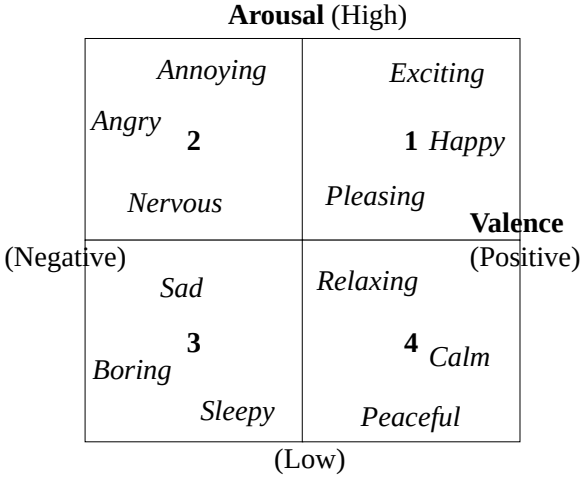


FIGURE 4.1: Circumplex model of emotion (Russell, 1980).

In order to identify all these different human emotions in music, emotion-related features are used in Music Emotion Recognition (MER) tasks (Yang and Chen, 2011). In this thesis, we only need to compare emotions, but we do it by comparing these features, extracted from the clips selected from the content selection phase and from the music. Emotion-related music features are then presented and described in the following section.

4.2 Emotion-related Music Features

Emotions transmitted by music are highly associated with different patterns of acoustic cues (Hevner, 1935; Juslin, 2000; Krumhansl). Table 4.2 relates arousal and valence to musical cues.

TABLE 4.2: Arousal and valence associated to acoustic cues (Gabrielsson, 2001).

Arousal	Tempo (fast/slow)
	Pitch (high/low)
	Loudness (high/low)
	Timbre (bright/soft)
Valence	Mode (major/minor)
	Harmony (consonant/dissonant)

Emotion perception is almost never dependent on a single music factor but on a combination of music factors (Hevner, 1935; Rigg, 1964). For instance, loud chords and high-pitched chords may be related to more positive valence than soft chords and low-pitched chords, irrespective of mode.

If we consider energy, rhythm, temporal, spectrum, and harmony as perceptual dimensions of music listening, we can extract features that best represent them (Table 4.3).

TABLE 4.3: Extracted Feature Sets

Feature Set	**Features**
Energy	Dynamic loudness, audio power (AP), total loudness (TL), and specific loudness sensation coefficients
Rhythm	Beat histogram, rhythm pattern, rhythm histogram, and tempo, rhythm strength, rhythm regularity, rhythm clarity, average onset frequency, and average tempo (Lu et al., 2006)
Temporal	Zero-crossings, temporal centroid, and log attack time
Spectrum	Spectral centroid, spectral rolloff, spectral flux, SFM, and SCF, MFCC, spectral contrast (Jiang et al., 2002), DWCH (Li and Ogihara, 2006), tristimulus, even-harm, and odd-harm (Wieczorkowska, 2004), roughness, irregularity, and inharmonicity
Harmony	Salient pitch, chromagram centroid, key clarity, pitch histogram, SWIPE (Camacho, 2007)

Spectrum and temporal features summarize the timbre content of a song (Yang and Chen, 2011).

4.2.1 Energy Features

The energy of a song is frequently highly correlated with arousal perception (Gabrielsson, 2001).

We can measure perceived loudness by the dynamic loudness model of Chalupper and Fastl (2002), modeling parameters of auditory sensation based on some psychoacoustic models, such as the Bark critical ban (Zwicker, 1961) for modeling auditory filters in our ears, an auditory temporal integration model, and a Zwicker and Fastl (1999) model for modeling sharpness.

We can also extract audio power (AP), total loudness (TL), and specific loudness sensation coefficients (SONE), which are energy-related features as well. Experiences show that audio power (AP) and total loudness (TL) for each frame are related to the arousal perception of music pieces and the SONE coefficients of songs of different emotions may have different characteristics (Yang and Chen, 2011). AP is the power of the audio signal, while the extraction of TL and SONE is based on an outer-ear model, the Bark critical-band rate scale (psycho-acoustically motivated critical bands), and spectral masking (by applying spreading functions). The resulting power spectrum, the sonogram, represents better human loudness sensation. The sonogram consists of up to 24 Bark critical-bands, depending on the sampling frequency of the audio signal. SONE coefficients are computed from the sonogram based on the Hartmann (1997) method. TL is computed as an aggregation of SONE, which takes the sum of the largest SONE coefficient and a 0.15 ratio of the sum of the remainder coefficients. All these features can be extracted for each short time frame and then aggregated by taking the average and standard deviation for temporal integration (Meng et al., 2007).

4.2.2 Rhythm Features

Experiences show that rhythm features are associated with the perception of both valence and arousal (Yang and Chen, 2011).

Rhythm is the pattern of pulses/notes of varying strength and it is often described in terms of tempo, meter, or phrasing. A song with fast tempo commonly means that it has high arousal. Apart from that, flowing rhythm is normally correlated to positive valence, whereas firm rhythm is correlated to negative valence (Gabrielsson, 2001).

We can obtain some rhythm features from the beat histogram of music, including beat strength, amplitude and period of the first and second peaks of the beat histogram, and the ratio of the strength of the two peaks in terms of beats per minute (bpm). The beat histogram is constructed by computing the autocorrelation of the signal envelope in each octave frequency band. The dominant peaks of the autocorrelation function are various periodicities of the signal's envelope.

We can also extract rhythm patterns to see how strong and fast beats are played within a specific frequency band (Pampalk et al., 2002). We can apply short-time Fourier transform (STFT) to obtain the amplitude modulation of SONE of each segment of a music piece. Repetitive patterns in the individual modulation frequency indicates the presence of rhythm. The rhythm pattern of the entire music piece can be obtained calculating the median of the rhythm patterns of its segments (no overlapping).

Rhythm strength, rhythm regularity, rhythm clarity, average onset frequency, and average tempo are also relevant to both valence and arousal perception. Considering "onset" as the starting time of each musical event (note), rhythm strength is calculated as the average onset strength of the onset detection curve, that we can compute using the algorithm described by Klapuri (1999). Rhythm regularity and rhythm clarity can be acquired by performing autocorrelation on the onset detection curve. If a music segment has a regular rhythm, the peaks of the corresponding autocorrelation curve will be strong. Onset frequency corresponds to the number of note onsets per second, while tempo corresponds to the periodicity of the onset detection curve. We can also estimate tempo, obtaining the mean of a 60-bin rhythm histogram, which sums the amplitude modulation coefficients across critical bands.

4.2.3 Temporal Features

Zero-crossing rate, temporal centroid, and log attack time are useful temporal features we can extract from music.

Zero-crossing rate is a measure of the signal noisiness and is computed by taking the mean and standard deviation of the number of signal values that cross the zero axis

in each time window (Equation 4.1, where T is the length of the time window, s_t is the magnitude of the t-th time-domain sample, and $w(\Delta)$ is a rectangular window). The standard deviation of zero-crossing rate may be useful for valence prediction (Yang and Chen, 2011).

$$\text{zero-crossing rate} = \frac{1}{T}\sum_{t=m-T+1}^{m}\frac{|\text{sgn}(s_t) - \text{sgn}(s_{t-1})|}{2}w(m-t) \tag{4.1}$$

Zero-crossing rate is normally high for noise and speech, moderate for music with vocals, and low for instrumental music.

Temporal centroid and log attack time are two timbre descriptors that depict the energy envelope (Allamanche, 2001). Temporal centroid is the average time over the energy envelope, while log attack time is the logarithm of the duration between the time the signal starts (defined as the time the signal reaches 50% of the maximum energy value, by default) and the time the signal reaches its maximum energy value.

4.2.4 Spectrum Features

Spectrum features are features computed from the STFT of an audio signal (Peeters, 2004). One can extract the timbral texture features including spectral centroid, spectral rolloff, spectral flux, spectral flatness measures (SFM), and spectral crest factors (SCF). These features are extracted for each frame and then by taking the mean and standard deviation for each second. The sequence of feature vectors is then collapsed into a single vector representing the entire signal by taking again the mean and standard deviation (Tzanetakis and Cook, 2002).

The average spectral centroid is highly related to arousal perception (Yang and Chen, 2011). It consists of the center of gravity of the magnitude spectrum of STFT and is calculated using Equation 4.2, where A_n^t is the magnitude of the spectrum at the t-th frame and the n-th frequency bin, and N is the total number of bins. The centroid is a measure of the spectral shape. Higher spectral centroid means "brighter" audio texture.

$$\text{spectral centroid} = \frac{\sum_{n=1}^{N} nA_t^n}{\sum_{n=1}^{N} A_t^n} \tag{4.2}$$

The standard deviation of spectral rolloff seems to be correlated with arousal (Yang and Chen, 2011). It is defined as the frequency k_t below which a certain fraction of the total energy is contained (Equation 4.3).

$$\sum_{n=1}^{k_t} A_t^n = 0.85 * \sum_{n=1}^{N} A_t^n \tag{4.3}$$

Spectral rolloff is another measure of the spectral shape and estimates the amount of high frequency in the signal.

Spectral flux estimates the amount of local spectral change and is given by Equation 4.4 (Tzanetakis and Cook, 2002), where a denotes the magnitude of the spectrum, normalized for each frame.

$$\text{spectral flux} = \sum_{n=1}^{N} (a_t^n - a_{t-1}^n)^2 \tag{4.4}$$

Tonalness is often related to the valence perception: joyful and peaceful melodies are tonal (tone-like), while angry melodies are atonal (noise-like) (Thompson and Robitaille, 1992).

SFM and SCF can help describe the tonalness of audio signal (Allamanche, 2001). SFM is the ratio between the geometric and arithmetic average of the power spectrum (Equation 4.5), whereas SCF is the ratio between the peak amplitude and the root-mean-square amplitude (Equation 4.6).

$$\text{spectral flatness measure} = \frac{(\prod_{n \in B^k} A_t^n)^{1/N_k}}{\frac{1}{N_k} \sum_{n \in B^k} A_t^n} \tag{4.5}$$

$$\text{spectral crest factor} = \frac{\max_{n \in B^k} A_t^n}{\frac{1}{N_k} \sum_{n=1}^{N} A_t^n} \tag{4.6}$$

B^k denotes the k-th frequency subband, and N_k is the number of bins in B^k.

We can also extract MFCC, which are the coefficients of the discrete cosine transform (DCT) of each short-term log power spectrum defined on a nonlinear perceptual-related Mel-frequency scale (Davis and Mermelstein, 1980), to represent the formant peaks of the spectrum.

Octave-based spectral contrast considers the spectral peak, spectral valley, and their dynamics in each subband to describe, although roughly, the relative distribution of the harmonic and non-harmonic components in the spectrum (Jiang et al., 2002).

Daubechies wavelets coefficient histogram (DWCH) features have better ability to represent both local and global information of the spectrum than traditional features, due to the use of the wavelet technique (Li and Ogihara, 2006). DWCH is computed from the Daubechies wavelet coefficients at different frequency subbands with different resolutions (Li and Ogihara, 2004, 2003, 2006).

We can also generate three sensory dissonance features, namely roughness, irregularity, inharmonicity.

Roughness seems to be correlated with our valence perception (Yang and Chen, 2011). It measures the noisiness of the spectrum, based on the fact that any note that does not respect prevailing harmony is considered dissonant. It is estimated by computing the peaks of the spectrum and taking the average of all the dissonance between all possible pairs of peaks (Sethares, 2005).

Irregularity measures the degree of variation of successive peaks of the spectrum (Fujinaga and MacMillan, 2000) and it is computed using Equation 4.7, where the square of the difference of the amplitude of adjoining partials is summed (Jensen, 1999).

$$\text{irregularity} = \frac{\sum_{n=1}^{N}(A_t^n - A_t^{n+1})^2}{\sum_{n=1}^{N} A_t^n * A_t^n} \tag{4.7}$$

Inharmonicity (Equation 4.8, where f_n is the n-th harmonic of fundamental frequency (f_0)) is computed as an energy-weighted divergence of the spectral components from the multiples of the f_0 (Peeters, 2004).

$$\text{inharmonicity} = \frac{2}{f_0}\frac{\sum_{n=1}^{N}|f_n - nf_0|(A_t^n)^2}{\sum_{n=1}^{N}(A_t^n)^2} \tag{4.8}$$

The inharmonicity represents the divergence of the signal spectral components from a strictly harmonic signal. Its value ranges from 0 (harmonic) to 1 (inharmonic).

Tristimulus measures the mixture of harmonics. The first tristimulus (Equation 4.9) measures the relative weight of the first harmonic; the second tristimulus (Equation 4.10) measures the relative weight of the second, third, and fourth together; and the third tristimulus (Equation 4.11) measures the relative weight of the remaining.

$$\text{tristimulus1} = \frac{(A_t^1)^2}{\sum_{n=1}^{N}(A_t^n)^2} \tag{4.9}$$

$$\text{tristimulus2} = \frac{\sum_{n=2,3,4}(A_t^n)^2}{\sum_{n=1}^{N}(A_t^n)^2} \tag{4.10}$$

$$\text{tristimulus3} = \frac{\sum_{n=5}^{N}(A_t^1)^2}{\sum_{n=1}^{N}(A_t^n)^2} \tag{4.11}$$

Even-harm and odd-harm represent the even and odd harmonics of the spectrum (Wieczorkowska, 2004; Wieczorkowska et al., 2005, 2006) (Equation 4.12 and Equation 4.13).

$$\text{even-harm} = \sqrt{\frac{\sum_{n=1}^{N/2}(A_t^{2n})^2}{\sum_{n=1}^{N}(A_t^n)^2}} \tag{4.12}$$

$$\text{odd-harm} = \sqrt{\frac{\sum_{n=1}^{N/2+1}(A_t^{2n-1})^2}{\sum_{n=1}^{N}(A_t^n)^2}} \tag{4.13}$$

4.2.5 Harmony Features

Harmony features are features computed from the sinusoidal harmonic modeling of the signal (Peeters, 2004). Musical sounds are harmonic, meaning that each sound consists of a series of multiplied frequencies over the lowest frequency, f_0.

Harmony features include two pitch features: salient pitch and chromagram center. A chromagram is the CQT for a vector containing the added complex module of the bins that correspond to octaves. Experiences show that high arousal values are usually associated with high average pitch and positive valence values seem to be associated with higher standard deviation of pitch values (Yang and Chen, 2011). The pitch (the perceived f_0) of each short time frame is estimated based on the multi-pitch detection algorithm described by Tolonen and Karjalainen (2000). The algorithm decomposes an audio waveform into two frequency bands (below and above 1 kHz), computes the autocorrelation function of the envelope in each subband, and estimates pitch by picking the peaks from the sum of the two autocorrelation functions. The highest peak's pitch estimate is the salient pitch.

One can also compute the wrapped chromagram for each frame and use the centroid of the chromagram as another estimate of f_0. This feature is called the chromagram centroid. A wrapped chromagram projects the frequency spectrum onto 12 bins that represent the 12 distinct semitones/chroma of the musical octave (it does not consider absolute frequency).

Harmony features also include three tonality features: key clarity, mode, and harmonic change. Each bin of the chromagram corresponds to one of the twelve semitone classes in the Western twelve-tone equal temperament scale. By comparing a chromagram to the 24 major and minor key profiles (Gómez, 2006), we can perform key detection and estimate the strength of the frame in association with each key. The highest key strength is returned as the key clarity.

Mode is often related to the valence perception of music (Gabrielsson, 2001; Oliveira and Cardoso, 2009) and describes a certain fixed arrangement of the diatonic tones of an octave (Oliveira and Cardoso, 2008).

We can use the algorithm developed by Harte et al. (2006) to compute a 6-dimensional feature vector called tonal centroid from the chromagram and use it to detect the harmonic changes, such as chord change, in musical audio, and then aggregate them by taking mean and standard deviation (Fastl, 1982).

We can also generate features like tonic, main pitch class, octave range of the dominant pitch, main tonal interval relation, and the overall pitch strength by computing the pitch histogram (Tzanetakis and Cook, 2002). Additionally, we can compare 16 pitch-related features including the mean, standard deviation, skewness, and kurtosis of the pitch and pitch strength time series estimated by Sawtooth Waveform Inspired Pitch Estimator (SWIPE), and SWIPE' (Camacho, 2007).

4.3 Summary

In this chapter, we presented well-known approaches to emotion characterization, such as categorical-based and dimension-based. Additionally, we described emotion-related music features used in MER, including those used in this thesis to address emotion coherence of our final artifact, which are: intensity features (computation of root-mean-square signal frame energy), timbre features (computation of zero-crossing rate of time signal, the voicing probability computed from the Autocorrelation Function (ACF), f_0 computed from the Cepstrum, and spectral features), and rhythm features (computation of Fast Fourier Transform (FFT) coefficients). These are detailed in the next chapter.

Chapter 5

Generation Of A Movie Tribute

This chapter presents our approach for the generation of a movie tribute.

We first proceed to data preprocessing, which includes subtitles' text, and music's and movies' sound normalization.

Considering a movie tribute to be a videoclip that contains clips from a movie playing with a song in background, we face the problem of collecting the most important parts of the movie to make the tribute credible. In order to determine the film's most important content, and avoid diversity (to guarantee coherence), we summarize its subtitles using the centrality-based LexRank algorithm, and use the timestamps to obtain the corresponding video clips.

To assure emotional coherence among the selected clips, we focus on the music's emotion-related features. So, we extract emotion-related audio features from the music and the scenes the clips belong to, and compare them to obtain the scenes that are emotionally more similar to the music. Emotion-related audio features include intensity, timbre, and rhythm features.

The final video is composed by joining the top-ranked clips that are also in the selected scenes from the emotional coherence addressing phase, and the input music.

Post-production (volume adjustment) is necessary in order to avoid conflicts between the clips' audio stream and the music's.

Figure 5.1 shows an overall scheme of the production of a movie tribute. Figure 5.2 presents a more detailed diagram containing the processes involved.

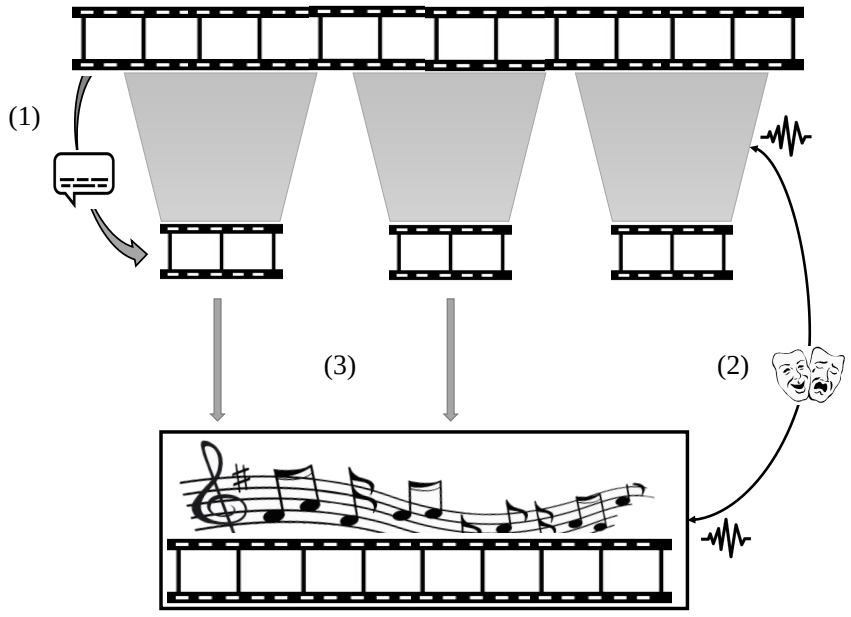


FIGURE 5.1: Movie tribute generation. (1) content selection. (2) emotion synchronization. (3) video composition.

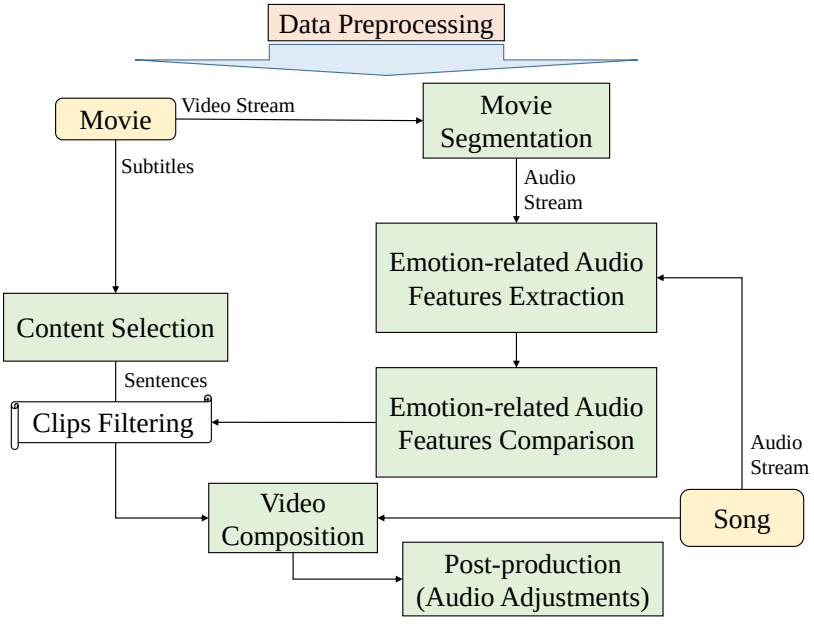


FIGURE 5.2: Movie tribute generation processes.

5.1 Data Preprocessing

All subtitles were segmented at the sentence level. Timestamps and punctuation inside sentences were removed.

We normalize the volume of the music pieces and the movies to a standard value to mitigate the production effect (some songs are recorded with a higher volume, while others are recorded with a lower volume). Our approach to volume normalization is to look for the loudest volume of the audio waveform and then amplify or attenuate the entire waveform until the loudest volume reaches 0 dB (peak normalization). We perform it using Ffmpeg (Bellard et al., 2003). We also trim the songs in order to remove silence at the beginning and ending of each song, using Wavosaur, a free audio editing software.

5.2 Content Selection

We use subtitles on account of their total coverage of the movies' dialogs, which can give us narrative information, easy text-audio/video stream mapping (using their timestamps), and availability.

In order to maintain some consistency concerning the selected content, we use a centrality-based approach. For this reason, we obtain relevant sentences from the movie, summarizing its subtitles, using LexRank. We chose LexRank due to previous work (Aparício et al., 2015) that shows that LexRank is the centrality-based algorithm to provide better summaries using the movie's subtitles, according to Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004). The length of the summary corresponds to the number of seconds of the song given as input.

In LexRank's implementation, the damping factor is set to 0.85 and the convergence threshold to 0.0001. After the similarity between each pair of sentences is calculated, each sentence's score is iteratively updated by the algorithm until there is convergence. Finally, sentences are picked until summary imposed length is reached.

5.3 Emotional Coherence

For each clip obtained from the content selection phase, we detect the scene in which it appears. Then, we extract emotion-related audio features from the music and the scenes of the video clips, and compare them to obtain the ones that are more emotionally similar to the music. We chose to compare the clip's scene instead of the clip itself for the scenes may provide more auditory information concerning the events involving the clip.

5.3.1 Movie Segmentation

In order to obtain the movie's scenes, we segment the movie's video stream using Lav2yuv, a program distributed with the MJPEG tools (Chen et al., 2012).

5.3.2 Feature Extraction

All music pieces are converted to a unique format: 22,050 Hz sampling frequency, 16 bits precision, and mono channel.

5.3.2.1 Intensity Features

The intensity feature set is represented by the average and standard deviation of the root-mean-square signal frame energy, extracted with OpenSMILE (Eyben et al., 2013).

5.3.2.2 Timbre Features

In order to characterize the music's timbre, we calculate the average and standard deviation of the following features (50 ms frames, no overlap): MFCC (1-12), zero-crossing rate of time signal (frame-based), the voicing probability computed from the ACF, f_0 computed from the Cepstrum, spectral centroid, spectral spread, spectral skewness,

spectral kurtosis, spectral flatness, spectral flux, spectral rolloff, spectral brightness, spectral entropy.

These features were also extracted with OpenSMILE (Eyben et al., 2013).

5.3.2.3 Rhythm Features

We consider high frequencies 9-dimensional rhythmic features and low frequencies 9-dimensional rhythmic features to represent rhythm (Antunes et al., 2014). Low frequencies rhythmic features are computed from FFT coefficients on the 20 Hz to 100 Hz range and high frequencies on the 8000 Hz to 11025 Hz range.

Considering v as a matrix of FFT coefficients with frequency represented by its columns and time by its lines, each component of the 9-dimensional vector is: $maxamp$ (maximum of the average v along time), $minamp$ (minimum of the average v along time), number of v values above 80% of $maxamp$, number of v values above 15% of $maxamp$, number of v values above $maxamp$, number of v values below $minamp$, mean distance between peaks, standard deviation of distance between peaks, maximum distance between peaks.

These features were extracted using a Matlab implementation by Antunes et al. (2014).

5.3.3 Feature Comparison

The resulting vector for each clip is compared with the music's vector using the cosine distance (Equation 5.1). If the similarity between them is greater than 0.7 (empirically determined value), we consider that the video clip has the same emotion of the music.

$$\text{cosine similarity}(A, B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{5.1}$$

A and B are two different vectors.

5.3.4 Video Composition

The length of the audio clip containing the music is filled with the clips resulting from the content selection phase, following the chronological order of the input movie. In the end, a black screen appears containing "Thank you for the memories, [movie's main character][1]".

5.3.4.1 Subtitles Timestamps Adjustments

The text stream is mapped to the video using subtitles, occasionally causing the time interval corresponding to the sentences of the subtitles not to encompass the speech that it is portraying. To resolve the identified abrupt shot transitions, the underlying audio stream is used to provide continuity cues. For that, a data-drive voice activity detector based on Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) is used (Eyben et al., 2013) to extend and/or compress the subtitles' timestamps.

5.4 Post-production: Volume Adjustments

We use OpenSMILE (Eyben et al., 2013) to obtain the energy produced throughout the music piece to determine the moments where it presents a higher volume, which interferes with the voices from the video. The music's volume is reduced to 30% whenever its energy is higher than -7.82 (logarithmic energy). We also extract the movie's energy with the same intent (but to reduce its own volume), so that audio continuity is not broken with high variances of loudness from one clip to another (due to the movie's soundtrack playing loud, for instance).

In order to retrieve sound energy from music, we extract prosodic features, which include f_0, the voicing probability, and the loudness contours. f_0 is computed via the sub-harmonic sampling algorithm. Pitch smoothing is done with a modification of the Viterbi algorithm (Ryan and Nudd, 1993).

[1] Given as input.

5.5 Summary

In this chapter, we presented our tribute generation method, including details about data preprocessing, the content selection method, the way we addressed emotional coherence (including the movie segmentation tool we used, which features were extracted, and how they were compared), final video composition (including subtitles timestamps adjustments), and post-production (volume adjustments).

Chapter 6

Experiments

In this chapter, we present the dataset used in our experiments, and information on the subjects that evaluated the final tributes, followed by the evaluation and the discussion of the results.

6.1 Dataset

Five tributes were generated using five different movies and songs (Table 6.1)

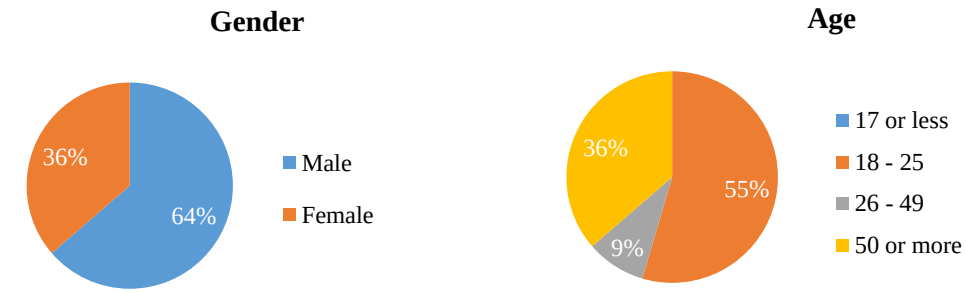
TABLE 6.1: Generated tributes.

Movie	**Song**	**Duration**
"Atonement" (2007)	"La Plage" (by Yann Tiersen)	02:01
"300" (2006)	"To The Edge" (by Lacuna Coil)	03:18
"Furious 7" (2015)	"See You Again" (by Wiz Khalifa)	03:57
"The Curious Case of Benjamin Button" (2008)	"The Last Goodbye" (by Billy Boyd)	04:13
"Interstellar" (2014)	"Conspiracy Agent" (by Savant)	01:01

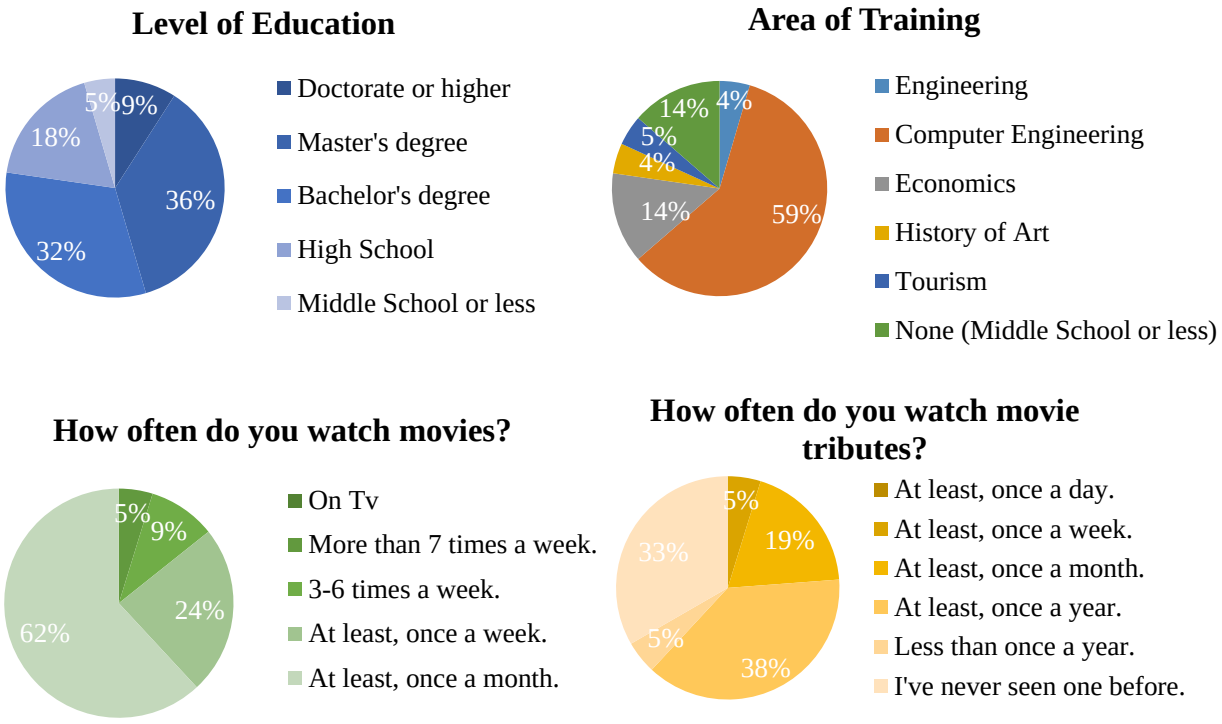
"To The Edge", "See You Again", and "The Last Goodbye" contain vocals, while the remaining don't.

6.2 Setup

22 subjects were invited to participate in this experiment. Figure 6.1 shows a characterization of the viewers, answering the following questions: "What is your age?", "What is your gender?", "What is your level of Education?"; "What is your area of training?"; "How often do you watch movies?"; "How often do you watch movie tributes?".



(A)



(B)

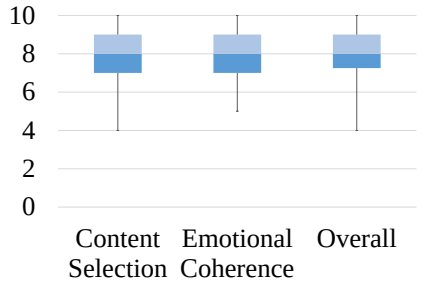
FIGURE 6.1: Viewers characterization.

Most people were males (64%), had between 18 and 25 years (55%), had a bachelor's degree (32%) or a master's degree (36%), and were Computer Engineers (59%).

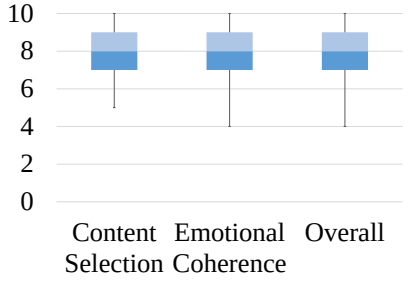
Most people watches movies once a month, at least (62%), and 33% didn't know what a movie tribute was, while 38% watches one once a year, at least.

6.3 Results

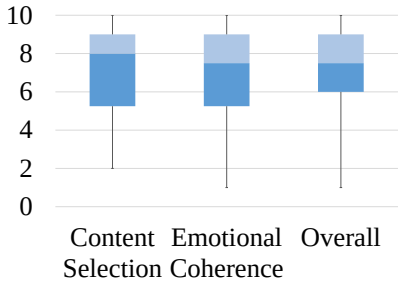
For each tribute, we present the results for content selection, and emotional coherence criteria, and overall scores (Figure 6.2).



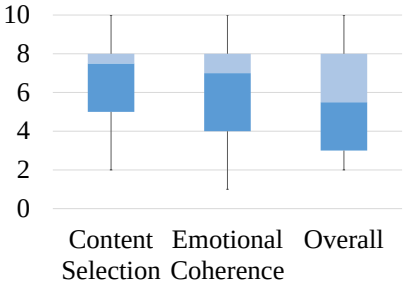
(A) "The Curious Case of Benjamin Button"



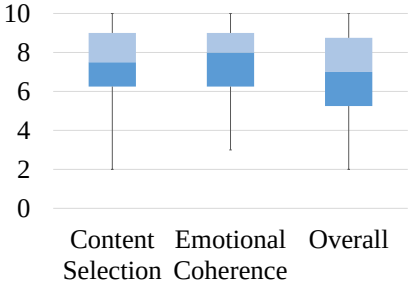
(B) "Atonement"



(C) "Interstellar"



(D) "300"



(E) "Furious"

FIGURE 6.2: Each tribute's evaluation results.

The tributes that had the best scores were “The Curious Case of Benjamin Button” and “Atonement”, with an average of 8 points in both content selection, and emotional coherence criteria, and overall evaluation, on a scale from 1 to 10.

“300” obtained the worst scores, with less than 6 points on overall evaluation, despite having 7.5 on content selection criteria and 6.9 on emotional coherence criteria.

All tributes obtained more than 7 points on content selection and more than 6.9 point in emotional coherence. On overall evaluation, the minimum of points given were 5.5. On average, our method led to scores above 7 on both content selection, and emotional coherence criteria, and overall evaluation (Figure 6.3).

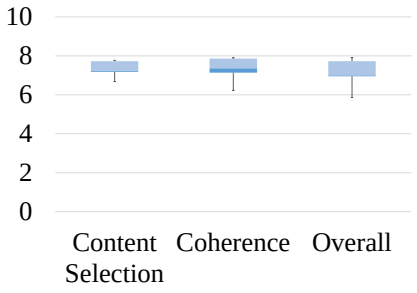
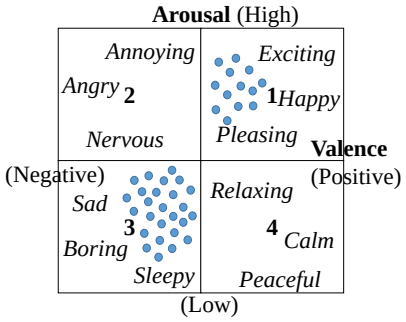


FIGURE 6.3: Average tributes’ evaluation results.

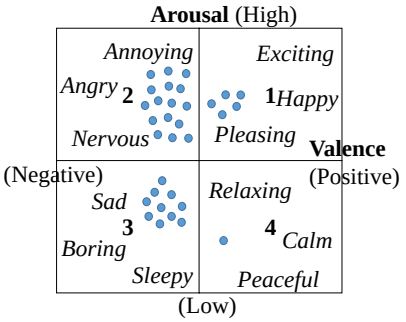
Regarding content selection, while some evaluators thought the selected content was adequate, it was sometimes suggested the inclusion of specific scenes, or more important ones, or even later parts of the movie, in the final tribute. “300”, “Furious 7” and “The Curious Case of Benjamin Button” tributes were considered too big by some viewers. A few evaluators considered that the end of the tributes ended abruptly, and a few others considered it negative to appear similar scenes in different clips.

Concerning emotional coherence, the clips have not always been considered consistent with each other and hence with the music. It was sometimes expected that the most intense portions of the music (e.g., the chorus) would show more aggressive scenes, when all of them illustrated very calm scenes, suggesting the scenes could match the pace of the music, or even be rearranged to fit better with the music parts. It was highly pointed out the music was not well chosen, in the case of “300”, and “Furious 7”.

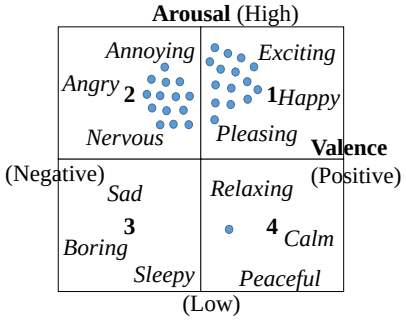
One of the main post-production related critics had to do with the overlap of song/-dialogs, suggesting the speech volume should be higher than the background music. Some disliked the music of the tribute being mixed with the original soundtrack and actors speaking. In addition, there are some abrupt transitions, so that some clips should continue when the song is in a steady move (in a verse, for instance), avoiding forced breaks. Some clips ended with speech being interrupted. It was also suggested that contiguous clips of the same scene should not exist.



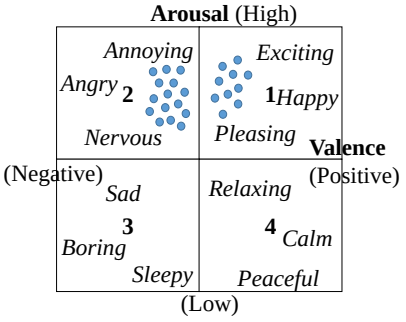
(A) "The Curious Case of Benjamin Button"



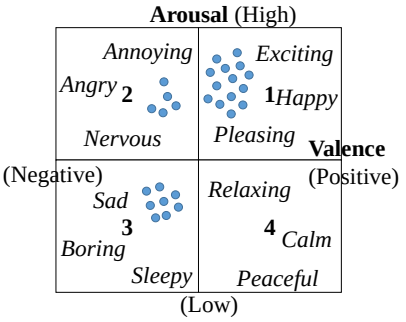
(B) "Atonement"



(C) "Interstellar"



(D) "300"



(E) "Furious 7"

FIGURE 6.4: Each tribute's identified emotions distribution in the circumplex model of emotion (Russell, 1980).

The identified emotions for each tribute (Figure 6.4) were diverse and belonged to different areas in the Russell's circumplex model of affect (Russell, 1980). The "The Curious Case of Benjamin Button" and "Furious 7" tributes are the most emotionally diverse, while the others tend to show emotions that are present in a certain pole of either valence or arousal ("Atonement" lays on negative valence, and "Interstellar" and "300" lay on high arousal).

An average of 23% of viewers might have given less than 7 points on tribute overall evaluation because they hadn't watched the movie yet (and watch movies on a regular basis, at least, once a month; which might mean they don't like this type of movie) and/or don't/didn't like the song, not because they thought the tribute was poorly produced.

An average of 42% of viewers who gave 7 points or more on tribute overall evaluation had already watched the movie, which means the tribute may have achieved its purpose: to relive the movie in a quick, effective and pleasant way.

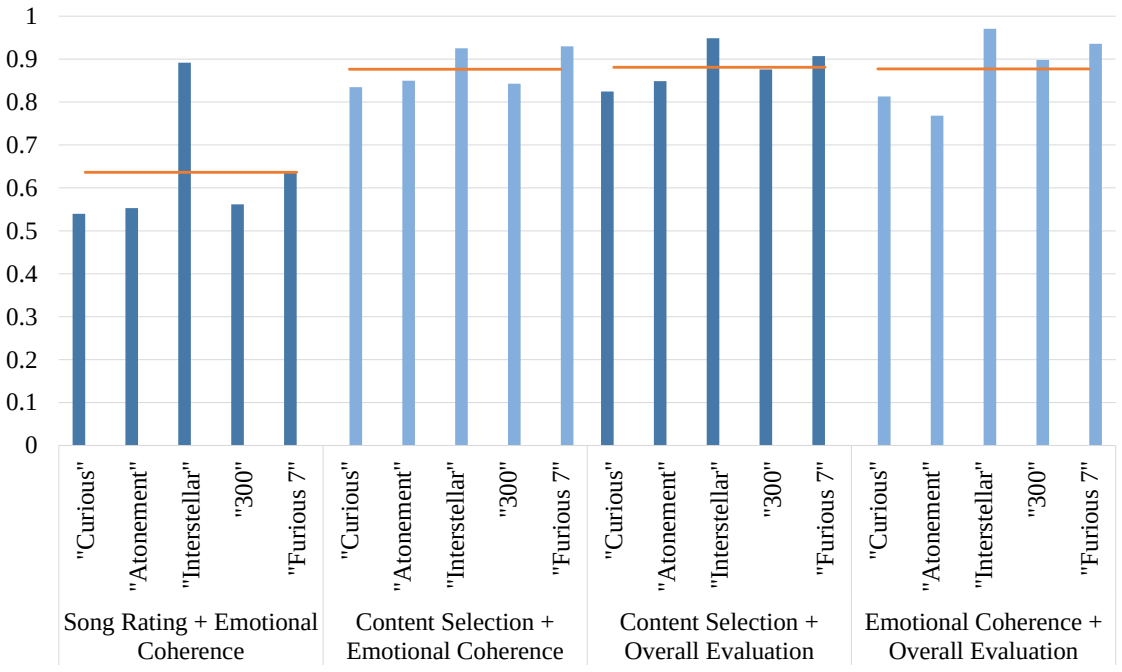


FIGURE 6.5: Spearman's ranks.

Figure 6.5 shows that content selection and emotional coherence convey strong, positive correlation, as well as content selection and emotional coherence, and overall

scores. In fact, content selection seems to affect the tribute's emotional coherence and content selection and emotional coherence seem to be directly related to the overall scores. On the other hand, except for "Interstellar", the song's appreciation doesn't seem to influence the tribute's emotional coherence.

Despite the critics, some viewers were astonished by the produced tributes, for knowing they were generated automatically. Some were interested in watching more of this kind of videos on YouTube.

6.4 Discussion

In this work, content selection is driven by a text stream that corresponds to transcripts of speech monologues and dialogs, presented in the input document's subtitles. In that sense, important content is not detected based on visual or audio cues, except those corresponding to speech (via subtitles). Hereof, other approaches can be used, using audio and visual cues for video abstraction (Coldefy and Bouthemy, 2004).

For text-based selection, different approaches are available depending on the aspects of interest (such as diversity). For movie tributes, which target the viewer's emotions, algorithms that focus on the most central (important) content may be indicated. Apart from LexRank, other algorithms can be used, for instance, Support Sets, which is also a centrality-based algorithm.

Subtitles are segmented at sentence-level. The song is used to obtain an emotionally-coherent multimedia artifact. Results show that segmentation at sentence-level does not affect, significantly, its overall coherence.

6.5 Summary

This chapter presented the results and an analysis of the produced tributes, and ended with an overall discussion.

Chapter 7

Conclusions

This thesis presented methods for the generation of multimedia artifacts, specifically, movie tributes, focusing on content selection and coherence aspects.

Considering a movie tribute to be a videoclip that contains clips from a movie playing with a song in background, we faced the problem of collecting the most important parts of the movie to make a tribute. In order to determine the film's most important content, we summarize its subtitles using the LexRank algorithm and use the timestamps to obtain the corresponding video clips.

To guarantee coherence among the selected clips, we focus on the music's emotion-related features. So, we extract audio features that describe emotions from the music and the scenes of the clips, and compare them to obtain the scenes more similar to the music.

The final video is composed by joining the top-ranked clips that are also in the selected scenes from the previous phase, and the input music with adjusted volume.

Five tributes were produced and the human evaluation was positive, having, on average, achieved scores above 7 (on a scale from 1 to 10). All tributes obtained more than 7 points on content selection and more than 6.9 point in emotional coherence. On overall evaluation, the minimum of points given were 5.5.

Content selection seems to affect the tribute's emotional coherence and content selection and emotional coherence seem to be directly related to the overall scores.

Ours results show that segmentation at sentence-level does not affect, significantly, the overall coherence.

7.1 Contributions

This thesis offers a simple and tested way to generate multimedia artifacts, specifically, videoclips that portray a film with the aid of music, combining these two distinct and valuable cultural areas. It provides a means to remember films in just a few minutes, as well as promote them and eventually raise its number of visualizations.

Two papers were produced and submitted to the arXiv. The first one was "Summarization of Films and Documentaries Based on Subtitles and Scripts", and here we assess the performance of generic summarization algorithms when applied to subtitles and scripts, for films and documentaries. The second one was "Generation of Multimedia Artifacts: An Extractive Summarization-based Approach", where we explore methods for content selection and address the issue of coherence in the context of the generation of multimedia artifacts.

7.2 Future Work

Regarding future work and concerning the identified issues, if the music has vocals, its lyrics can be taken into account to relate their topics to the movie. A possible solution is to receive only the movie as input, then, choose a topic-related song from an existing dataset, for instance, The Million Song Dataset (Bertin-Mahieux et al., 2011). Furthermore, to improve the final artifact's structural coherence, we can take into account the music's structure and align it to the video stream (Nieto and Bello, 2014).

Still regarding coherence, locally-coherent sentences for the summarized movie can be identified. LSA can be used as a technique for measuring coherence, by comparing vectors of adjacent sentences in the generated semantic space. Thus, they can be considered as groups of locally-coherent sentences.

Coherence can also be established by means of composition techniques for video production, based on temporal constraints, along with thematic and structural continuity (Ahanger and Little, 1998). In fact, these can be used as a means to build a narrative, which can be seen as a series of events in a chain (Branigan, 1992). Ahanger and Little (1998) establishes content progression in the final multimedia artifact by comparing adjacent video segments and ensuring that they are not too similar or too different.

Bibliography

Gulrukh Ahanger and Thomas D. C. Little. Automatic composition techniques for video production. *IEEE Trans. Knowl. Data Eng.*, 10(6):967–987, 1998.

Eric Allamanche. Content-based identification of audio material using mpeg-7 low level description. In *ISMIR*, 2001.

Arnon Amir, Gal Ashour, and Savitha Srinivasan. Automatic gen. of conf. video proc. *J. Vis. Comun. Img. Represent.*, 15(3):467–488, 2004.

Keith Anderson and Peter W. McOwan. A real-time automated system for the recognition of human facial expressions. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 36(1):96–105, 2006.

Pedro Girão Antunes, David Martins de Matos, Ricardo Ribeiro, and Isabel Trancoso. Automatic Fado Music Classification. *CoRR*, 2014.

Marta Aparício, Paulo Figueiredo, Francisco Raposo, David Martins de Matos, and Ricardo Ribeiro. Summarization of Films and Documentaries Based on Subtitles and Scripts. *CoRR*, abs/1506.01273, 2015.

Fabrice Bellard, Michael Niedermayer, Juan J Sierralta Pulento, et al. Ffmpeg multimedia system, 2003. URL `http://ffmpeg.sourceforge.net/`. accessed 2015/09/21.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proc. of the 12th Intl. Conf. on Music Inf. Retr. (ISMIR 2011)*, 2011.

Stefano Bocconi, Frank Nack, and Lynda Hardman. Automatic generation of matter-of-opinion video documentaries. *Web Semant.*, 6(2):139–150, 2008.

Christoph Brachmann, Hashim I. Chunpir, S. Gennies, B. Haller, Thorsten Hermes, Otthein Herzog, Arne Jacobs, Philipp Kehl, Astrid P. Mochtarram, Daniel Möhlmann, Christian Schrumpf, C. Schultz, B. Stolper, and Benjamin Walther-Franks. Automatic generation of movie trailers using ontologies. *IMAGE - J. of Interdisciplinary Image Science*, 5:117–139, 2007.

Edward Branigan. *Narrative Comprehension and Film*. Routledge, 1992.

Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. of the 7th Intl. Conf. on World Wide Web*, pages 107–117, 1998.

Arturo Camacho. *Swipe: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. PhD thesis, Gainesville, FL, USA, 2007. AAI3300722.

Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proc. of the 21st Annual Intl. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr.*, pages 335–336, 1998.

J. Chalupper and H. Fastl. Dynamic loudness model for normal and hearing-impaired listeners. 88:378–386, 2002.

Junzhou Chen, Qing Li, Qiang Peng, and Kin Hong Wong. CSIFT based locality-constrained linear coding for image classification. *CoRR*, abs/1309.7484, 2013.

Lei Chen, Narasimha Shashidhar, and Qingzhong Liu. Scalable secure mjpeg video streaming. In *AINA Workshops*, pages 111–115, 2012.

M. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. Technical report, Carnegie Mellon University, Pittsburgh, USA, 2009.

F. Coldefy and P. Bouthemy. Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis. In *Proc. ACM Multimedia 2004*, New-York, October 2004.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech and Signal Proc.*, 28(4):357–366, 1980.

Duo Ding, Florian Metze, Shourabh Rawat, Peter Franz Schulam, Susanne Burger, Ehsan Younessian, Lei Bao, Michael G. Christel, and Alexander Hauptmann. Beyond audio and video retrieval: Towards multimedia summarization. In *Proc. of the 2Nd ACM Intl. Conf. on Multimed. Retr.*, pages 2:1–2:8, 2012.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, pages 169–200, 1992.

Günes Erkan and Dragomir R. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *J. of Artif. Intell. Res.*, pages 457–479, 2004.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. on Multimed.*, pages 1553–1568, 2013.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent Dev.s in openSMILE, the Munich Open-source Multimed. Feature Extractor. In *Proc. of the 21st ACM Intl. Conf. on Multimed.*, pages 835–838, 2013.

H. Fastl. Fluctuation strength and temporal masking patterns of amplitude-modulated broad-band noise. *Hearing Research*, 8(1):59–69, 1982.

Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Sci.*, pages 972–976, 2007.

I. Fujinaga and K. MacMillan. Realtime recognition of orchestral instruments. In *Proc. of the Intl. Comput. Music Conf.*, volume 141, page 143, 2000.

A. Gabrielsson. The Influence of Musical Structure on Emotional Expression. *Music and Emotion: Theory and Research*, 2001.

E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, 2006.

Yihong Gong and Xin Liu. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proc. of the 24th Annual Intl. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr.*, pages 19–25, 2001.

Christopher Harte, Mark Sandler, and Martin Gasser. Detecting harmonic change in musical audio. In *Proc. of the 1st ACM Workshop on Audio and Music Comput. Multim.*, pages 21–26. ACM, 2006.

W.M. Hartmann. *Signals, Sound, and Sensation*. Modern Acoustics and Signal Proc. American Inst. of Physics, 1997.

K. Hevner. Expression in music: a discussion of experimental studies and theories. *Psychological Review*, 42(2):186–204, 1935.

Xian-Sheng HUA, Lie LU, and Hong-Jiang ZHANG. Automatic music video generation based on temporal pattern analysis. In *Proc. of the 12th Annual ACM Intl. Conf. on Multimed.*, pages 472–475. ACM, 2004.

Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. Optimization-based automated home video editing system. *IEEE Trans. on Circuits and Systems for Video Technology*, 14 (5):572–583, 2004a.

Go Irie, Kota Hidaka, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Latent topic driving model for movie affective scene classification. In *ACM Multimed.*, pages 565–568, 2009.

Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Automatic trailer generation. In *Proc. of the Intl. Conf. on Multimed.*, pages 839–842, 2010.

L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Inf. Process. Systems, Vol. 19 (NIPS*2005)*, pages 547–554, 2006.

K. Jensen. *Timbre Models of Musical Sounds: From the Model of One Sound to the Model of One Instrument*. Københavns Universitet, Datalogisk Institut, 1999.

Dan-ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast features. In *Multim. and Expo, 2002. ICME '02. Proc. 2002 IEEE Intl. Conf. on*, volume 1, pages 113–116, 2002.

Patrik N Juslin. Cue utilization in communication of emotion in music performance: Relating performance to perception. *J. Experimental Psychology: Human Perception and Performance*, 26(6):1797–1813, 2000.

P.N. Juslin and J. Sloboda. *Handbook of Music and Emotion: Theory, Research, Applications*. OUP Oxford, 2011.

A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Acoustics, Speech, and Signal Proc., 1999. Proc., 1999 IEEE Intl. Conf. on*, volume 6, pages 3089–3092 vol.6, 1999.

Peter Kolb. Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*, volume 4, pages 81–88. Northern European Association for Language Technology, 2009.

C. Krumhansl. Music: A link between cognition and emotion. *Current Directions Psychological Sci.*, 11(2):45–50.

Thomas K Landauer and Susan T. Dutnais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Process.*, 25(2-3):259–284, 1998.

T. Li and M. Ogihara. Detecting Emotion in Music. October 2003.

Tao Li and M. Ogihara. Content-based music similarity search and emotion detection. In *Acoustics, Speech, and Signal Proc., 2004. Proc. (ICASSP '04). IEEE Intl. Conf. on*, volume 5, pages V–705–8, 2004.

Tao Li and M. Ogihara. Toward intelligent music information retrieval. *IEEE Trans. on Multim.*, 8(3):564–574, 2006.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summ. Branches Out: Proc. of the ACL-04 Workshop*, pages 74–81, 2004.

Lie Lu, D. Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *Trans. on Audio Speech and Lang. Process.*, 14(1):5–18, 2006.

Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proc. of the 10th ACM Intl. Conf. on Multimed.*, pages 533–542. ACM, 2002a.

Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the Tenth ACM International Conference on Multimedia*, pages 533–542. ACM, 2002b.

P. Melville, S. M. Yang, M. Saar-Tsechansky, and Raymond J. Mooney. Active learning for probability estimation using jensen-shannon divergence. In *Proc. of the 16th European Conf. on Mach. Learn.*, pages 268–279, 2005.

Engin Mendi, Hélio B. Clemente, and Coskun Bayrak. Sports video summarization based on motion analysis. *Comput. Electr. Eng.*, 39(3):790–796, 2013.

A. Meng, P. Ahrendt, J. Larsen, and L.K. Hansen. Temporal feature integration for music genre classification. *Audio, Speech, and Lang. Proc., IEEE Trans. on*, 15(5): 1654–1664, 2007.

George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11): 39–41, 1995.

Gabriel Murray, Steve Renals, and Jean Carletta. Extractive Summarization of Meeting Recordings. In *Proc. of the 9th European Conf. on Speech Commun. and Technology*, pages 593–596, 2005.

Tomoyasu Nakano, Sora Murofushi, Masataka Goto, and Shigeo Morishima. Dancereproducer: An automatic mashup music video generation system by reusing dance video clips on the web. In *Proc. of the 8th Sound and Music Comput. Conf. (SMC 2011)*, pages 183–189, 2011.

Oriol Nieto and Juan Pablo Bello. Music segment similarity using 2d fourier magnitude coefficients. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Process.*, pages 664–668, 2014.

NIST. 2011 trecvid multimedia event detection track. `http://www.nist.gov/itl/iad/mig/med11.cfm`, 2011. Accessed: January 28, 2012.

António Oliveira and Amilcar Cardoso. Controling music affective content: A symbolic approach. In *Conf. on Interdisciplinary Musicology*, 2008.

António Oliveira and Amilcar Cardoso. Automatic manipulation of music to express desired emotions. In *Sound and Music Comput.*, 2009.

Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. In *Proc. of the 10th ACM Intl. Conf. on Multim.*, pages 570–579. ACM, 2002.

G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, IRCAM, 2004.

Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

Rosalind W. Picard, E. Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(10):1175–1191, 2001.

R. Plutchik and H. Kellerman. *Theories of Emotion*. Elsevier Sci., 2013.

Ricardo Ribeiro and David Martins de Matos. Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. *J. of Artif. Intell. Res.*, pages 275–308, 2011.

M. G. Rigg. The mood effects of music: A comparison of data from four investigators. *J. Psychology*, 58:427–438, 1964.

J.A. Russell. A circumplex model of affect. *J. of Personality and Social Psychology*, 39 (6):1161–1178, 1980.

Matthew S. Ryan and Graham R. Nudd. The viterbi algorithm. Technical report, University of Warwick, 1993.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Intl. Conf. on New Methods in Lang. Process.*, pages 44–49, 1994.

Emery Schubert. Update of the hevner adjective checklist. *Perceptual and motor skills*, 96(3 Pt 2):1117–1122, 2003.

Björn Schuller, Clemens Hage, Dagmar Schuller, and Gerhard Rigoll. 'mister d.j., cheer me up!': Musical and textual features for automatic mood classification. *J. of New Music Res.*, 39(1):13–34, 2010.

William A. Sethares. *Tuning, timbre, spectrum, scale*. Springer, 2nd ed edition, 2005.

W. F. Thompson and B. Robitaille. Can composers express emotions through music? *Empirical Studies of the Arts*, 10:79–89, 1992.

T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. *Speech and Audio Proc., IEEE Trans. on*, 8(6):708–716, 2000.

G. Tzanetakis and P. Cook. Musical Genre Classification of Audio Signals. *IEEE Trans. on Speech and Audio Process.*, pages 293–302, 2002.

Ju-Chiang Wang, Yi-Hsuan Yang, Hsin-Min Wang, and Shyh-Kang Jeng. The acoustic emotion gaussians model for emotion-based music annotation and retrieval. In *Proc. of the 20th ACM Intl. Conf. on Multimed.*, pages 89–98, 2012.

Alicja Wieczorkowska. Towards extracting emotions from music. In Leonard Bolc, Zbigniew Michalewicz, and Toyoaki Nishida, editors, *IMTCI*, volume 3490, pages 228–238. Springer, 2004.

Alicja Wieczorkowska, Piotr Synak, Rory Lewis, and Zbigniew W.Raś. Extracting emotions from music data. In *Foundations of Intelligent Systems*, volume 3488, pages 456–465. Springer Berlin Heidelberg, 2005.

Alicja Wieczorkowska, Piotr Synak, and Zbigniew W. Ra's. Multi-label classification of emotions in music, 2006.

Xixuan Wu, Bing Xu, Yu Qiao, and Xiaoou Tang. Automatic music video generation: cross matching of music and image. In *Proc. of the 20th ACM Multimed. Conf., MM '12, Nara, Japan, October 29 - November 02, 2012*, pages 1381–1382, 2012.

Shasha Xie and Yang Liu. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *Proc. - ICASSP, IEEE Intl. Conf. on Acoustics, Speech and Signal Process.*, pages 4985–4988, 2008.

Y.H. Yang and H.H. Chen. *Music Emotion Recognition*. CRC Press, 2011.

Xiaojin Zhu, Andrew B. Goldberg, Jurgen Van Gael, and David Andrzejewski. Improving Diversity in Ranking using Absorbing Random Walks. In *Proc. of the 5th NAACL - HLT*, pages 97–104, 2007.

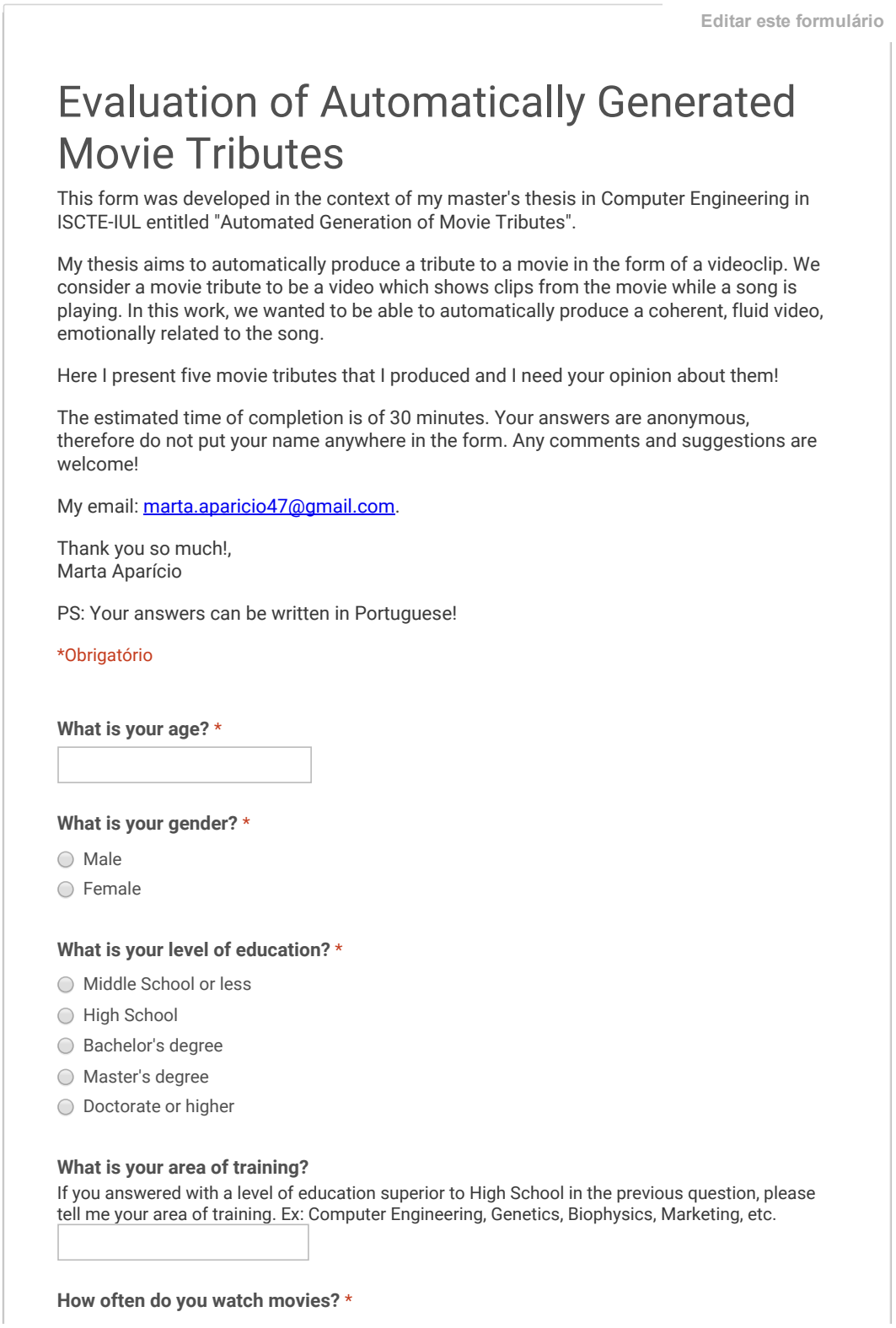
E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The J. of the Acoustical Society of America*, 33(2):248, 1961.

Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and Models (Springer Series in Information Sciences) (v. 22)*. Springer, 2nd updated edition, 1999.

Appendices

Appendix A

A User Test Interface



Editar este formulário

Evaluation of Automatically Generated Movie Tributes

This form was developed in the context of my master's thesis in Computer Engineering in ISCTE-IUL entitled "Automated Generation of Movie Tributes".

My thesis aims to automatically produce a tribute to a movie in the form of a videoclip. We consider a movie tribute to be a video which shows clips from the movie while a song is playing. In this work, we wanted to be able to automatically produce a coherent, fluid video, emotionally related to the song.

Here I present five movie tributes that I produced and I need your opinion about them!

The estimated time of completion is of 30 minutes. Your answers are anonymous, therefore do not put your name anywhere in the form. Any comments and suggestions are welcome!

My email: marta.aparicio47@gmail.com.

Thank you so much!,
Marta Aparício

PS: Your answers can be written in Portuguese!

*Obrigatório

What is your age? *

What is your gender? *

- Male
- Female

What is your level of education? *

- Middle School or less
- High School
- Bachelor's degree
- Master's degree
- Doctorate or higher

What is your area of training?
If you answered with a level of education superior to High School in the previous question, please tell me your area of training. Ex: Computer Engineering, Genetics, Biophysics, Marketing, etc.

How often do you watch movies? *

FIGURE A.1: General user information request (page 1).

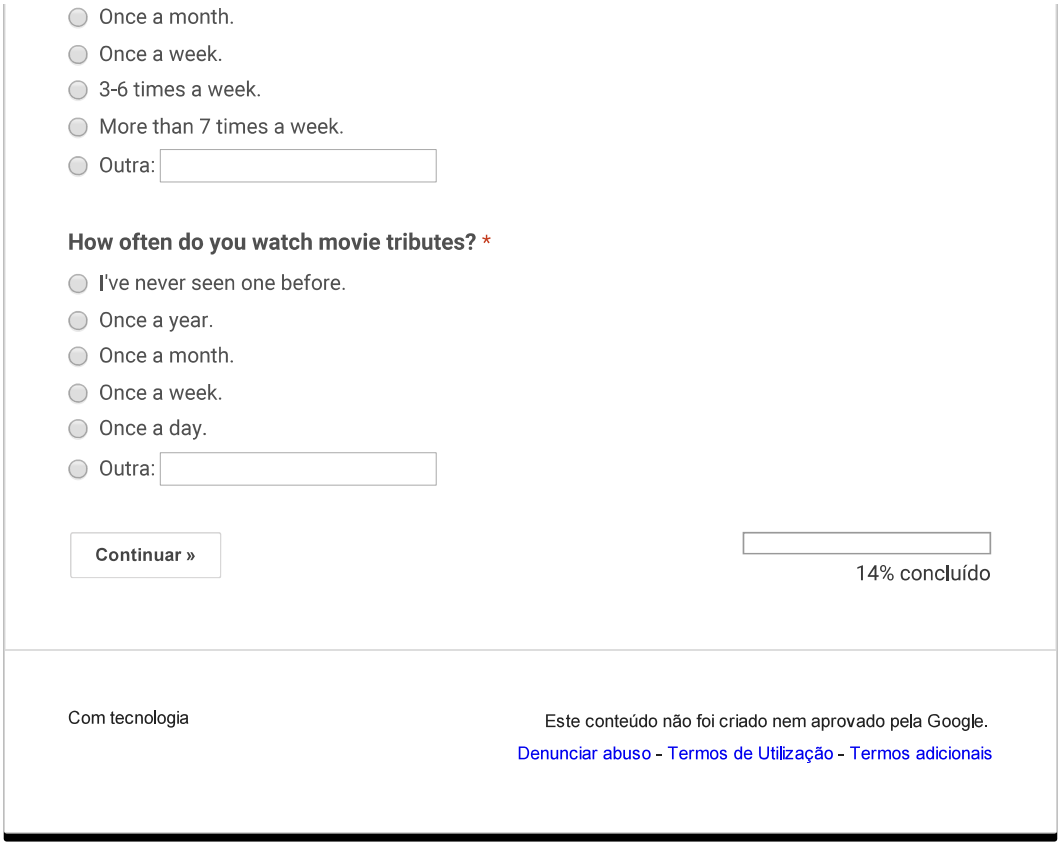


FIGURE A.2: General user information request (page 2).

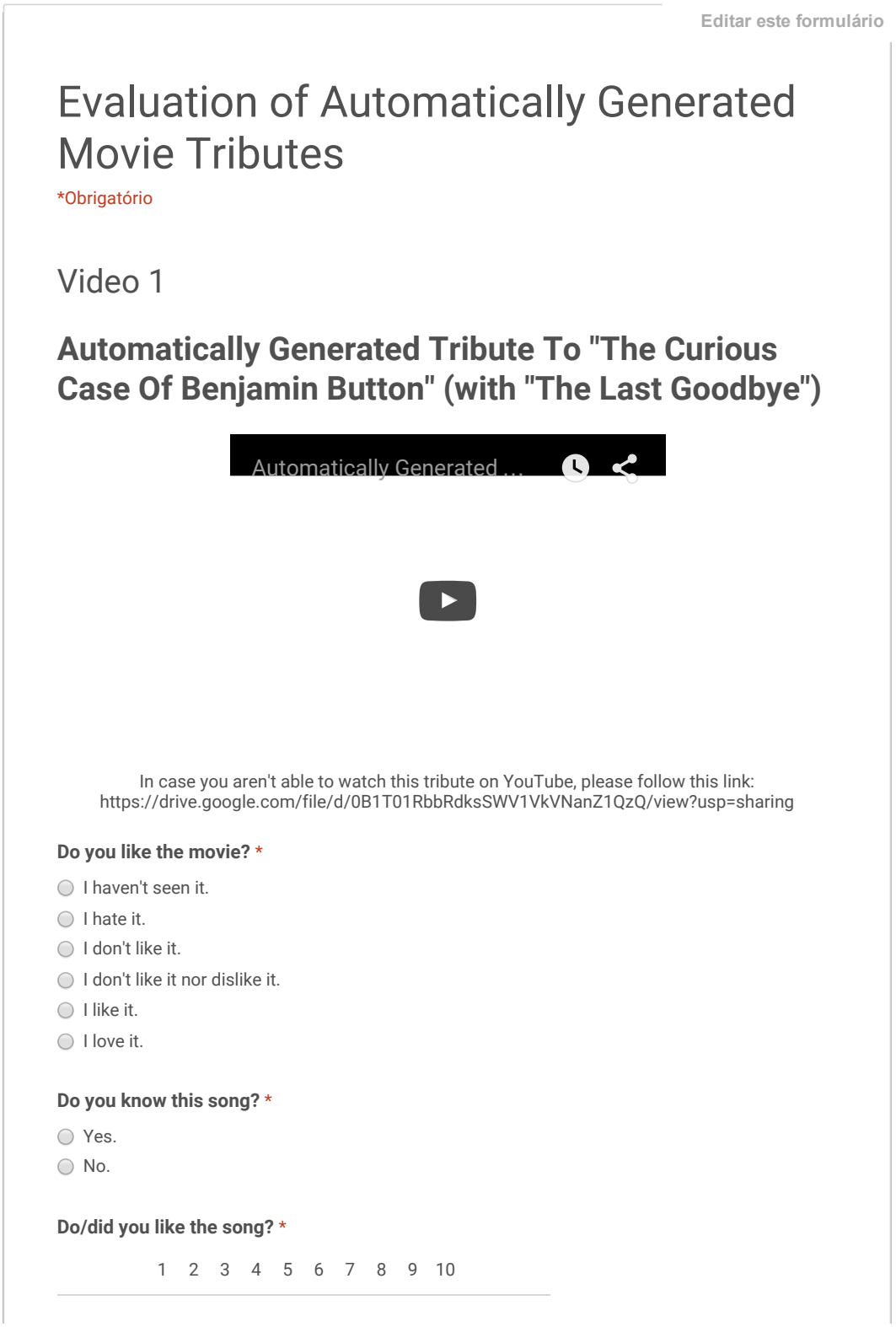


FIGURE A.3: Individual tribute evaluation (page 1).

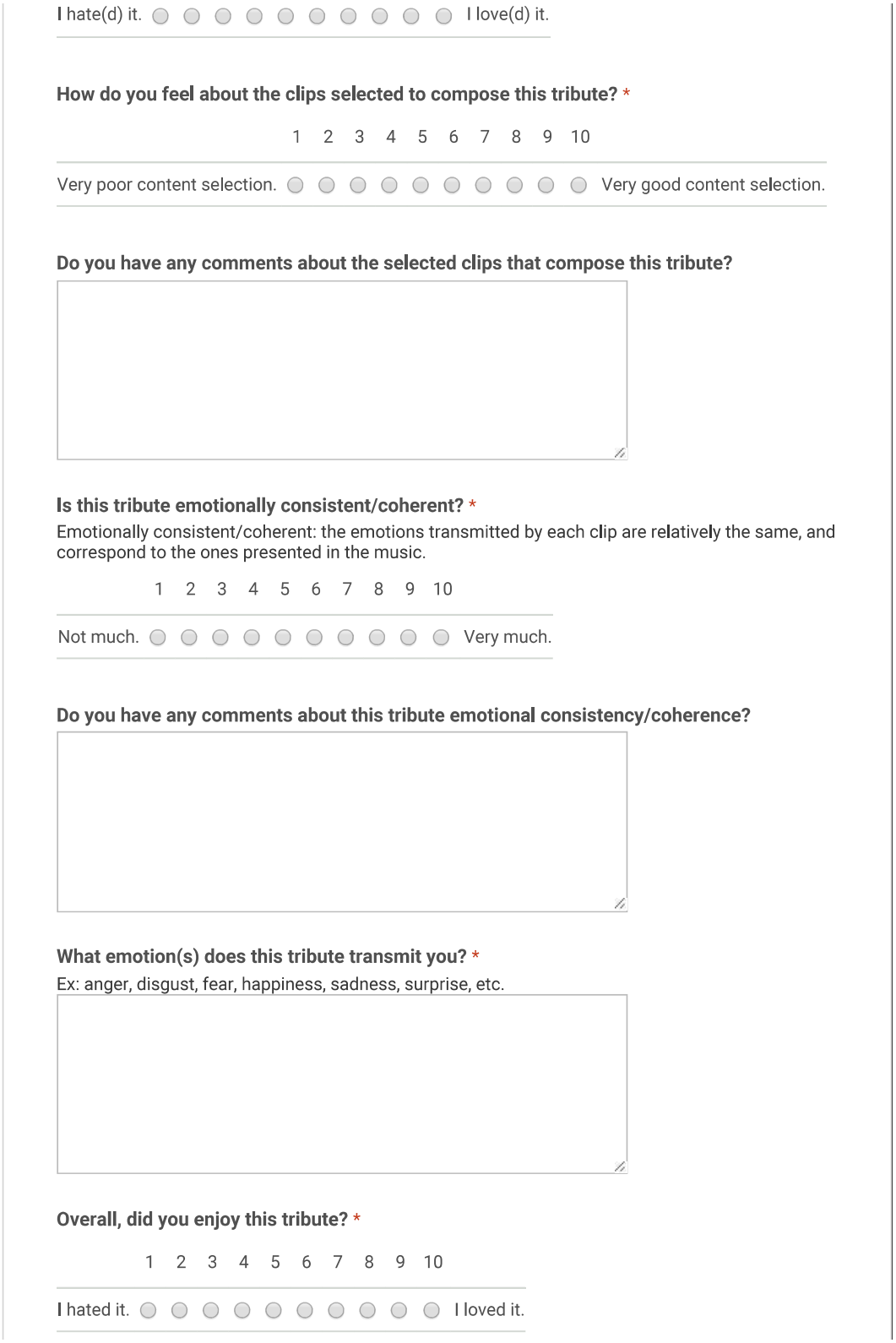


FIGURE A.4: Individual tribute evaluation (page 2).

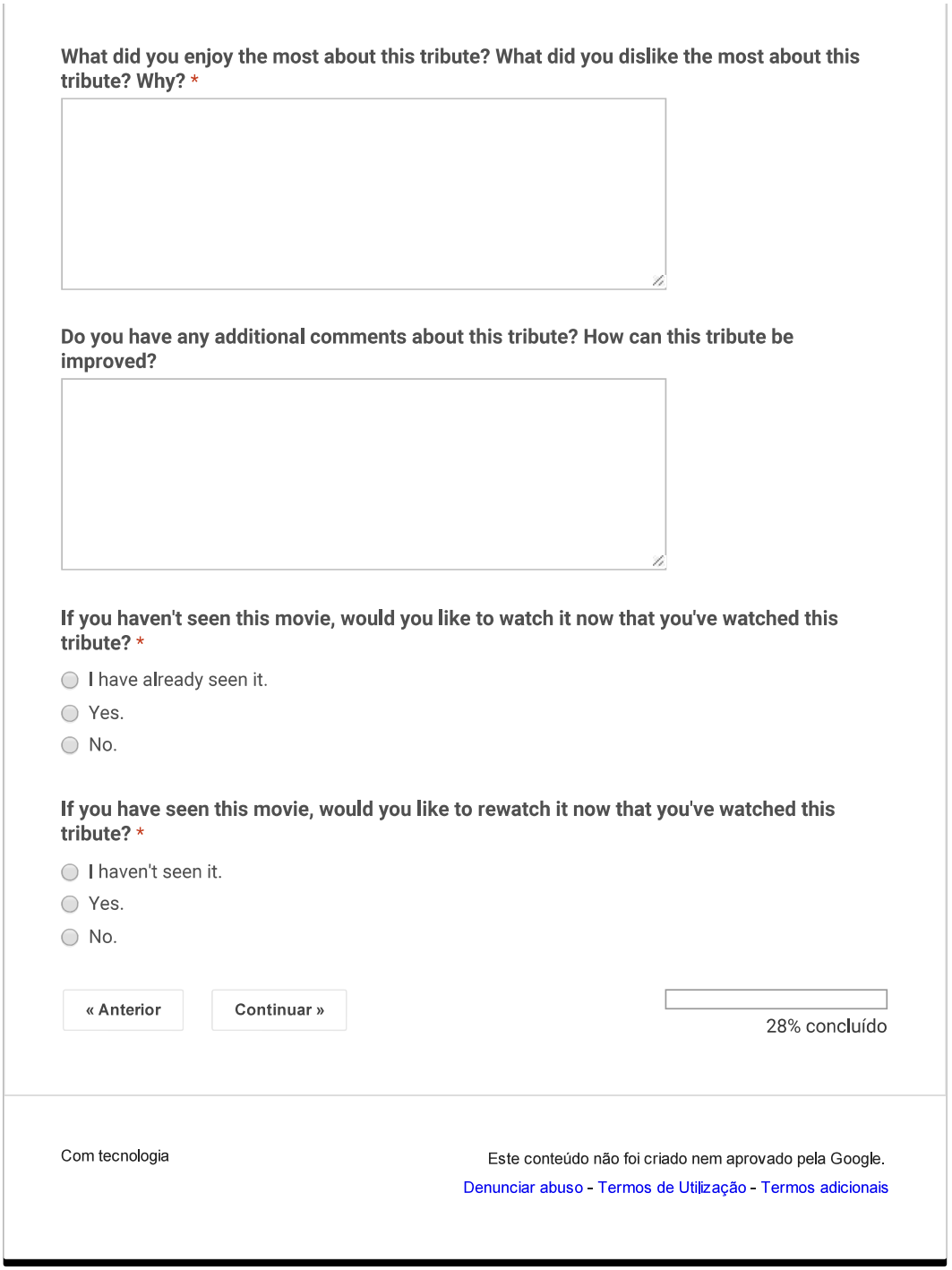


FIGURE A.5: Individual tribute evaluation (page 3).