The Influence of Audio Quality on the Popularity of Music Videos: A YouTube Case Study

Michael Schoeffler
International Audio Laboratories Erlangen*
Am Wolfsmantel 33
91058 Erlangen
michael.schoeffler@audiolabs-erlangen.de

Jürgen Herre
International Audio Laboratories Erlangen*
Am Wolfsmantel 33
91058 Erlangen
juergen.herre@audiolabs-erlangen.de

ABSTRACT

Video-sharing websites like YouTube contain many music videos. On such websites, the audio quality of these music videos can differ from poor to very good since the content is uploaded by users. The results of a previous study indicated that music videos are very popular in general among the users. This paper addresses the question whether the audio quality of music videos has an influence on user ratings. A generic system for measuring the audio quality on video-sharing websites is described. The system has been implemented and was deployed for evaluating the relationship between audio quality and video ratings on YouTube. The analysis of the results indicate that, contrary to popular expectation, the audio quality of music videos has surprisingly little influence on its appreciation by the YouTube user.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing

Keywords

Audio Quality; Popularity; Overall Listening Experience

1. INTRODUCTION AND RELATED WORK

Nowadays, music plays an essential role in the life of many people and has become pervasive in their everyday lifes [9, 5]. One reason for listening to music is to enjoy the performances and the listening experience [13, 1]. The question arises what are the factors which influence whether a particular piece of music is liked or not? Why music is liked or not liked is investigated by researchers from various areas but we are far away from having identified all factors [4].

In multimedia-related communities, the "overall satisfaction/enjoyment" of an user is called *Quality of Experience* [8]. In the context of listening to music, Schoeffler et al. introduced the term *Overall Listening Experience* [10]. Schoeffler et al. carried out two experiments about rating the Overall Listening Experience of music excerpts which had different levels of audio quality [11, 10]. The results of both experiments showed that ratings of the Overall Listening Experience are influenced by the audio quality. However, the effect size of the audio quality strongly depends on the individual listener [12].

Video-sharing websites with user-generated content (e. g. YouTube [15]) have videos in very different levels of audio quality. There are many possible reasons for degraded audio quality on such platforms since many technologies and processes are involved until a video is uploaded and can be watched by other users. It starts with the file format of the video file that is stored on the user's hard disk before uploading. The audio stream of video files is often encoded in a lossy compressed format like MP3, Vorbis or AAC. These compressed formats allow to encode the audio stream in different levels of audio quality depending on how much bit rate is spent. Even if users upload the video in a lossless compressed format, most video-sharing websites would not have enough storage space to store all videos. Furthermore, offering videos in a lossless compressed format to users who want to watch them will result in a very high demand of bandwidth. However, even though the provided audio quality is not excellent for all videos, there are indications that music-related videos are very popular on video-sharing websites. An analysis of the YouTube website traffic by Gill et al. [2] revealed that the music category is one of the most popular categories. One could assume, based on the popularity of the music category, that video-sharing websites like YouTube are also used for just listening to music without paying much attention to the video content. In the analysis by Gill et al. it was also found out that videos were generally rated high (the mean rating of videos in the most popular lists was consistently near 4 out of 5 with very little variation). These high ratings indicate that a lot of content is very popular among the users.

This paper addresses the research question whether the audio quality of such videos has an influence on the user ratings on YouTube. Conducting the study in such an uncontrolled environment is fully intended even if there are many possible factors which may influence the user ratings. This is, however, the condition under which music videos are usually consumed and is thus ecologically more valid than experiments under controlled lab conditions. By comparing the results of this study with results obtained from controlled experiments in the same context, a statement can be made to which degree controlled experiments reflect real world data.

*A joint institution between the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

WISMM'14, November 7, 2014, Orlando, Florida, USA.

ACM 978-1-4503-3157-9/14/11...$15.00.
http://dx.doi.org/10.1145/2661714.2661725.

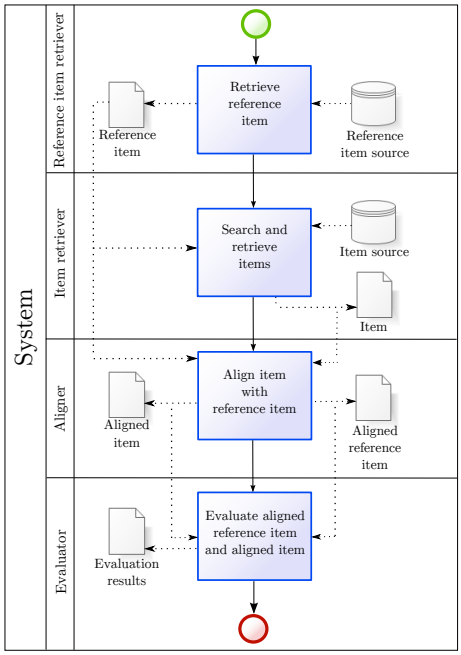

Figure 1: BPMN 2.0 diagram of the system.

2. SYSTEM

2.1 Design

The generic system contains three main functionalities: Retrieval, Alignment, Evaluation (see Figure 1). The audio quality evaluation is based on a comparison of the music video's audio stream and the original music signal having best-available audio quality. In the following, the audio stream of a music video is named *item* and the original music signal is named *reference item*. First of all, *reference items* are retrieved from a *reference item source* by the *reference item retriever*. The *reference item retriever* must retrieve at least the music signal and a term which describes the *reference signal*. A term could be the name of the performing artist and the name of the song. The *item retriever* retrieves *items* from an *item source* by using the *reference signals'*. The *item retriever* uses the terms of the *reference items* for searching corresponding *items* at the *item source*. An *item* must contain the music signal and values which represent the popularity of the music video. In preparation for the quality evaluation, the *aligner* aligns the music signals of the *items* and their corresponding *reference item*. Aligning the signals is especially required when the *reference item* or the *item* has only an excerpt of the music signal. The audio quality evaluation is applied by the *evaluator*. The *evaluator* takes an *item* and its corresponding *reference item* and compares both music signals with respect to their audio quality. As a result of the comparison, the *evaluator* returns one or more values which describe the differences in audio quality between the two music signals. By investigating the relationship between the audio quality values calculated by the *evaluator* and the popularity values of the *items*, a statement can be made whether audio quality has an influence on the popularity of music videos.

2.2 Implementation

2.2.1 Reference Item Retrieval

A reference item must be of original audio quality and has at least a duration of a few seconds. Google Play Store is an Internet store that sells, among other products, music as digital downloads [3]. For all available songs, Google Play Store offers free previews which are short music excerpts. These previews are 320 KBit/s CBR MP3s which are considered to have very high quality. All previews are in stereo and have a sample rate of 44100 Hz. The length of the previews is either 30 s or 90 s depending on the duration of the corresponding song. Web pages of the Google Play Store's music section present either a single release, an album release of an artist or a compilation of different songs from various artists. A typical web page of the music sections contains fifteen to twenty songs. The previews are used as reference items for the audio quality evaluation. Beside the preview, the name of the album, song and artist are collected. In addition, the duration of the song, the average rating (Five-Star Likert scale) and the number of ratings is acquired.

2.2.2 Item Retrieval

For each *reference item*, a search query is sent to YouTube to retrieve related music videos by using the YouTube API. The search query contains a search term according to which YouTube selects related music videos. The search term is chosen to be the reference item's artists name concatenated with the song name. The YouTube API responded to the search queries with lists containing a maximum of ten related music videos. A related music video of the list is only selected for further quality evaluation if three conditions are met:

- The duration of the music video must be similar to the *reference item*. The difference in duration must be less than eight seconds. Since the *reference items* have only excerpts, their duration was taken from the metadata which contained the duration of the complete song.
- Viewing the music video must not be restricted by country-specific restrictions.
- The title of the music video must not contain one of the following words: "live", "parody", "cover", "remix", "karaoke", "version", "instrumental" and "tutorial". If the search term contains one ore more of these words, the contained words are not covered by this condition.

The idea of these conditions is to heuristically exclude videos which do very likely not represent the reference item according to their title. Besides the audio stream of the music video, the YouTube category, duration, title, audio codec, average rating, number of ratings, number of views, number of likes, number of dislikes and number of marked as favorite video are collected. When we conducted a feasibility analysis for this project, we identified two types of music videos on YouTube. The first type of music videos contains a small film, the song video, clips of the performing artist or promotional material. This type of video is characterized by containing a lot of visual content. In contrast to such videos, we found some videos which just contain a single-colored background and showing the lyrics, a single pictures of the performing artist or a slide show of the artist. These videos have less visual information than the first type of videos. An influence of visual stimuli on the experience of music was confirmed by many researchers (e.g. Viollon et al. [14]) and should be considered in the statistical analysis of this work. Therefore, we additionally collect the size of the video per pixel as a prediction value of the information contained by a video:

$$videoSize_{\text{PerPixel}} = \frac{videoSize_{\text{total}}}{frameRate \cdot width \cdot height \cdot length}, \quad (1)$$

where $videoSize_{\text{total}}$ is the total file size of the video in byte, $frameRate$ the numbers of frames per second, $width$ the width of the video in pixel, $height$ the height of the video in pixel and $length$ the length of the video in seconds. The average size per pixel is taken since the more information

is contained by an encoded video, the more disk space is needed.

The audio streams of the music videos and additionally retrieved attributes are used as *items*.

2.2.3 Alignment

The reference items from Google Play Store are only excerpts of the items retrieved from YouTube but time-aligned signals are needed for the later audio quality evaluation. Two signals are time-aligned if they have the same starting point and the same duration. Before aligning the signals, all signals were resampled to 48 kHz and down-mixed to mono. A very basic technique, the cross-correlation, for time-aligning signals is applied:

$$(ref \star item)[n] = \sum_{i=1}^{n} ref[i] \cdot item[n+i], \tag{2}$$

where ref is the reference item and $item$ the item. The result of the cross-correlation is a vector of similarity values. The highest value of this vector indicates the alignment position where the two signals are the most similar. Since the result of the cross-calculation contains no information whether the two signals do represent the same song, a second alignment method is applied.

The second alignment method aligns the two signals based on their frequency domain representation. The spectrograms S_{item} and S_{ref} of the item and reference item are calculated. The magnitudes of the spectrograms' spectral column vectors are normalized by the euclidean norm. Then, S_{item} and S_{ref} are multiplied with each other to get the similarity matrix SM of the item and reference item. A vector of similarity values between item and reference item is retrieved by summing up the diagonals of SM. As for the cross-correlation, the highest value of this vector indicates the alignment position where the two signals are the most similar.

An alignment is successful if the position of the cross-correlation method and the position of the frequency domain method have a maximum deviation about the window size. The final position is the result of the cross-correlation, since the frequency domain method is not sample-accurate.

2.2.4 Quality Evaluation

All successfully aligned items and aligned reference items are evaluated by Perceptual Evaluation of Audio Quality (PEAQ) [6]. PEAQ compares a reference signal and a signal under test and stores the results in a vector of so-called Model Output Variables (MOVs). Each MOV represents a degradation type of the audio quality [6]. When all MOV values have been calculated, they are given into a neural network to produce the objective difference grade (ODG). The ODG is a "Mean Opinion Score"-like value which ranges from 0 ("imperceptible impairment") to -4 ("very annoying impairment"). Our system uses an implementation of the McGill University [7].

3. EXPERIMENT

3.1 Procedure

The reference item retriever crawled 5000 web pages on the Google Play Store and found 52752 reference items. Based on the reference items, the item retriever found 77245 items on YouTube (crawled from a location in Germany). The aligner was able to successfully align 61250 of these reference items and items. The evaluator managed to evaluate all aligned items with their corresponding aligned reference item resulting in 61250 evaluation results. The spectrograms of the frequency-domain alignment method was obtained using an FFTs with a window size of 8192 samples and an overlap of 4096 samples. The whole evaluation process lasted from November to December 2013.

3.2 Results

The music excerpt are well-liked on both web sites. On the Google Play Store, a reference item has an average reference item rating of 3.8 out of 5 stars (SD[1] $= 1.9$). On YouTube, an item has an average item rating of 4.9 out of 5 ($SD = 0.3$). The YouTube ratings were retrieved from the YouTube API which changed its rating system in March 2010. Since March 2010 YouTube has used a rating system that lets users indicate whether they like or dislike a video. However, until March 2010, YouTube used a 1-5 rating system in which 1 was the lowest rating that could be given. The average ratings contain the data from the newer and older rating scale. The average number of likes is 1085.4 ($SD = 10826.3$) and the average number of dislikes is 35.8 ($SD = 612.7$).

An analysis of the relationship between the audio quality and the popularity of the items has some pitfalls. First of all, videos on YouTube have very high ratings in average. Hence, the retrieved data set is sparse and skewed. E.g. 11271 reference items have no rating at all and 6415 items have less than three ratings. Furthermore, there are 7151 items which have only a single corresponding reference item and 26478 items have at least five corresponding reference items. Besides that, the audio quality evaluation is subject to a limitation. PEAQ, especially the objective difference grade, was designed to compare music signals with small impairments. The objective difference grade is produced by a neural network which was fitted to the results of listening tests, where audio codecs were evaluated by expert listeners. Music videos on YouTube sometime may have large impairments and YouTube users commonly are not expert listeners.

The item ratings represent, to some degree, the popularity of a music video. A linear regression model without interactions is calculated to investigate the item ratings in more detail. All items which have at least one item rating and one rating of their corresponding reference item were included in the linear regression model. The predictor variables are the reference item ratings, the objective difference grade, and the video size per pixel. The reference item ratings ($\beta = .042, t(49975) = 9.32, p = .000$), the objective difference grade ($\beta = .043, t(49975) = 9.50, p = .000$), and the video size per pixel ($\beta = -.046, t(49975) = -10.10, p = .000$) have all minor significant[2] effects on the item ratings. In addition, the model does not predict the item ratings very well ($\bar{R}^2 = .006$). To calculate a more meaningful linear regression model, a subset of the data set is selected which is more dense. Only samples are selected which have at least five item ratings, five reference item ratings and all reference items must have at least five corresponding items. By applying these constraints, 10003 samples remain in the data set. As in the linear regression model which was fitted with the complete data set, the reference item ratings ($\beta = .187, t(9999) = 19.0, p = .000$) have a stronger effect on the item ratings than the objective difference grades ($\beta = .034, t(9999) = 3.4, p = .000$). Compared to the effect size of the reference item ratings, the effect size of the video size per pixel is very low ($\beta = -.037, t(9999) = -3.7, p = .000$). As expected, the new linear regression model explains the data better ($\bar{R}^2 = .036$). Figure 2 shows a bivariate histogram of the subset. The bivariate histogram also indicates that audio quality had no influence on the item ratings.

In addition to the linear regression models, the relationship between the popularity of music videos and their au-

[1] M = mean, SD = standard deviation, β = standardized regression coefficient, $\bar{R}^2$ = Adjusted R-squared.

[2] The significance level α is defined as 0.05 in this paper.

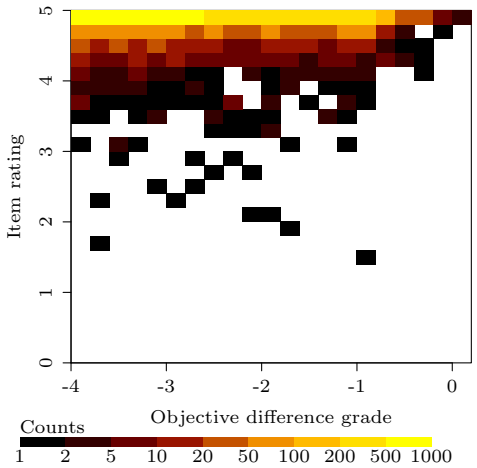

Figure 2: Bivariate histogram of the subset data (item ratings and audio quality of 10003 items).

dio quality is measured by Pearson's product-moment coefficient. The reference item ratings have a very weak correlation with item ratings ($r(10001) = .182, p = .000$). The objective difference grades have almost no correlation with item ratings ($r(10001) = .030, p = .003$). Table 1 contains more correlation values, including the Model Output Variables of PEAQ and the difference in average bandwidth between the reference item and item (AvgBandwidthDiff).

Variable	Person's r	p-value
Reference item rating	.182	.000
Video size per pixel	-.030	.003
Objective Difference Grade	.030	.003
$\text{WinModDiff}_\text{B}$	-.045	.000
$\text{AvgModDiff1}_\text{B}$	-.042	.000
$\text{AvgModDiff2}_\text{B}$	.011	.266
$\text{RmsNoiseLoud}_\text{B}$	-.072	.000
$\text{BandwidthRef}_\text{B}$	-.058	.000
$\text{BandwidthTest}_\text{B}$	-.002	.853
TotalNMR_B	-.045	.000
$\text{RelDistFrames}_\text{B}$	-.035	.000
MFPD_B	-.010	.320
ADB_B	-.061	.000
EHS_B	-.008	.446
AvgBandwidthDiff	-.024	.017

Table 1: Pearson's product-moment coefficient between various variables and item ratings.

3.3 Discussion

The analysis of the results confirms the findings of Gill et al. that music video content is very popular among YouTube-users[2]. To overcome the sparseness of the data set, more items could be retrieved by the item retriever. However, even if a more balanced and dense data set is retrieved, according to our results it is very unlikely that a reanalysis would reveal a noteworthy effect of the audio quality on the user ratings.

By investigating the relationship between the audio quality and the item ratings, no noticeable correlation has been found. As mentioned before in the Results section, the objective difference grade of PEAQ has not been optimized for evaluating the audio quality on video-sharing websites. Therefore, all MOVs provided by PEAQ were tested for correlation with the item ratings. However, none of the MOVs had a significant correlation with the item ratings. Such an outcome of the experiment was expected. On the one hand, the influence of audio quality on the overall listening experience could be clearly measured in controlled experiments [11, 10]. However, on the other hand, before conducting the experiment presented in this paper, a feasibility analysis was done where we handpicked some items and reference items and evaluated their quality difference by an informal subjective listening test. We were very surprised how distorted some music signals were but still received very good user ratings which indicated an absence of an association between audio quality and user ratings. Since the Overall Listening Experience is considered as a subset of the Quality of Experience, the mismatch between results retrieved from controlled experiments and retrieved from real world data does also concern Quality of Experience models as Many of these models are based on data retrieved from controlled experiments.

4. CONCLUSION

In this work a generic system was designed and implemented to investigate the relationship between the audio quality and user ratings of music on the video-sharing website YouTube. 61250 music videos were retrieved by the implemented system and evaluated in their audio quality. The statistical analysis of the results indicated that the audio quality has almost no effect on the user ratings. These results are in contrast to results achieved by controlled experiments where the influence of audio quality on the overall listening experience could be measured.

5. REFERENCES

[1] T. Chamorro-Premuzic and A. Furnham. Personality and music: can traits explain how people use music in everyday life? *British journal of psychology (London, England : 1953)*, 98(Pt 2):175–85, May 2007.

[2] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28, New York, NY, USA, 2007. ACM.

[3] Google Inc. "Google Play Store". `http://play.google.com`, Nov. 2013.

[4] A. E. Greasley and A. M. Lamont. Music preference in adulthood: Why do we like the music we do? In *9th International Conference on Music Perception and Cognition*, pages 960–966, Bologna, Italy, 2006.

[5] D. J. Hargreaves and A. C. North. The Functions of Music in Everyday Life: Redefining the Social in Music Psychology. *Psychology of Music*, 27(1):71–83, Apr. 1999.

[6] International Telecommunication Union. Rec. ITU-R BS.1387-1: Method for objective measurements of perceived audio quality, 2001.

[7] P. Kabal. An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality. *McGill University Technical Report*, 2002.

[8] P. Le Callet, S. Möller, and A. Perkis. Qualinet White Paper on Definitions of Quality of Experience (Version 1.1), 2012.

[9] A. C. North, D. J. Hargreaves, and J. J. Hargreaves. Uses of music in everyday life. *Music perception*, 22(1):41–77, 2004.

[10] M. Schoeffler, B. Edler, and J. Herre. How Much Does Audio Quality Influence Ratings of Overall Listening Experience? In *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 678–693, Marseille, France, 2013.

[11] M. Schoeffler and J. Herre. About the Impact of Audio Quality on Overall Listening Experience. In *Proceedings of Sound and Music Computing Conference 2013*, pages 48–53, Stockholm, Sweden, 2013.

[12] M. Schoeffler and J. Herre. About the Different Types of Listeners for Rating the Overall Listening Experience. In *Proceedings of Sound and Music Computing Conference 2014*, Athens, Greece, 2014.

[13] J. A. Sloboda, S. A. O'Neill, and A. Ivaldi. Functions of music in everyday life: An exploratory study using the experience sampling method. *Musicae scientiae*, 5(1):9–32, 2001.

[14] S. Viollon, C. Lavandier, and C. Drake. Influence of visual setting on sound ratings in an urban environment. *Applied Acoustics*, 63(5):493–511, May 2002.

[15] YouTube, LLC. Youtube. `http://www.youtube.com`, Nov. 2013.