Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality

Jianwei Shi  
jianwei.shi.cn@outlook.com  
L3S Research Center, Leibniz Universität Hannover

Christian Otto  
christian.otto@tib.eu  
Leibniz Information Centre for Science and Technology (TIB)

Anett Hoppe  
anett.hoppe@tib.eu  
Leibniz Information Centre for Science and Technology (TIB)

Peter Holtz  
p.holtz@iwm-tuebingen.de  
Leibniz Institut für Wissensmedien (IWM)

Ralph Ewerth  
ralph.ewerth@tib.eu  
Leibniz Information Centre for Science and Technology (TIB)

L3S Research Center, Leibniz Universität Hannover

ABSTRACT

Ranking and recommendation of multimedia content such as videos is usually realized with respect to the relevance to a user query. However, for lecture videos and MOOCs (Massive Open Online Courses) it is not only required to retrieve relevant videos, but particularly to find lecture videos of high quality that facilitate learning, for instance, independent of the video’s or speaker’s popularity. Thus, metadata about a lecture video’s quality are crucial features for learning contexts, e.g., lecture video recommendation in search as learning scenarios. In this paper, we investigate whether automatically extracted features are correlated to quality aspects of a video. A set of scholarly videos from a Mass Open Online Course (MOOC) is analyzed regarding audio, linguistic, and visual features. Furthermore, a set of cross-modal features is proposed which are derived by combining transcripts, audio, video, and slide content. A user study is conducted to investigate the correlations between the automatically collected features and human ratings of quality aspects of a lecture video. Finally, the impact of our features on the knowledge gain of the participants is discussed.

KEYWORDS

multimodal, video assessment, correlation, knowledge gain

ACM Reference Format:

Jianwei Shi, Christian Otto, Anett Hoppe, Peter Holtz, and Ralph Ewerth. 2019. Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality. In 1st International Workshop on Search as Learning with Multimedia Information (SALM '19), October 21, 2019, Nice, France. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3347451.3356731

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SALM ’19, October 21, 2019, Nice, France
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-6919-0/19/10...$15.00  
https://doi.org/10.1145/3347451.3356731

1 INTRODUCTION

Modern studying has left the classrooms – more and more, learners refer to the Web as a source of educational material for a certain learning need. This kind of informal learning scenario is, nowadays, only insufficiently supported by Web mechanisms such as search engines which focus on satisfying information needs instead ([2]). Recently, the research area Search as Learning (SaL) gains momentum: It recognizes learning as an implicit component of Web search [1], aims to detect learning needs and intents [10, 24] and to identify factors correlated to successful learning outcomes (e. g., [7]). While there are first interesting insights, the research field is still mostly geared to the recommendation of textual learning resources [17]. This stands in contrast to current learning research which suggests that users may prefer images and videos when tackling certain learning needs (e. g., procedural learning tasks [19, 25]). In consequence, this work sets out to establish a relationship between automatically extracted video features and successful learning. The objective is to facilitate more effective search for learning objects, be it for learners with a certain learning intent or for educators which seek for high-quality material to enhance their own courses. Normally, there is a large amount of educational and lecture videos that cover similar content – from single-video tutorials to full-fledged Massive Open Online Courses (MOOCs). In such a large lecture video archive, it is desirable to find content of high quality regarding the knowledge presentation. However, objective features which assess, for instance, the nature of the presenters voice or the design of the slides, are not fully explored in automated systems in the current state of research. A human viewer evaluates the quality of a learning resource based on all available information. In common lecture videos this includes the textual, oral and visual modality. Viewers are supported in their learning by the visual elements on the slide, the words spoken and the gestures of the lecturer.

Previous work, like Guo et al. [13], investigated how the design of lecture videos impacts the viewer engagement and provided recommendations to optimize the content accordingly. Chen et al. [6] used multimodal sensing to assess the quality of a presentation. They extracted speech, body movement and visual features from the shown slides. Principal Component Analysis was applied to human
ratings in order to address the two main modalities of the presentation: 1) recital skills, including, for instance, voice information and body language, and 2) slide quality, with regards to grammar, readability, and visual design. Pearson correlation was used to measure the relation between the different features. Haider et al. [14] proposed a system for automatic video quality assessment, which is the most similar to our approach, focusing on prosodic and visual features. They extracted the complete set of audio features from the ComParE challenge [22] and a total of 42 features related to hand movements of the speaker. The employed Multimodal Learning Analytics (MLA) dataset [18] contains 416 oral presentations (in Spanish) and the respective metadata regarding speech, facial expressions, skeletal data extracted from a Microsoft Kinect, as well as the shown slides. Each of these videos was labeled with ten individual ratings and an overall score related to the quality of the slides. A correlation study (discriminant analysis) was employed which found that prosodic features are able to predict self confidence and enthusiasm (of the speaker) as well as body language and pose, which is a quality measure their participants had to label. Their visual features showed similar results, but with less accuracy.

In this paper, we go beyond previous work by (1) proposing a novel set of intuitive unimodal and cross-modal features that do not rely on skeletal data which are hard to acquire, (2) conducting an empirical evaluation on the correlation of features with quality aspects of a video, and (3) conducting an empirical evaluation on the correlation of features with participants’ knowledge gain. Videos of a MOOC website are utilized and automatically annotated with our set of unimodal and multimodal features that address different quality aspects. Different modalities are exploited: audio and spoken text, video frame content, slide content, as well as cross-modal features. The experimental results reveal that several of our features show a significant correlation with the corresponding human assessment of the lecture videos.

The remainder of the paper is organized as follows. Section 2 describes the set of extracted unimodal as well as novel cross-modal features. The design of the user study and the experimental results are presented in Section 3, while Section 4 concludes the paper and outlines areas of future work.

## 2 EXTRACTION OF UNIMODAL AND CROSS-MODAL FEATURES

This section outlines our approach for the extraction of in total 22 features from lecture videos including textual, audio, linguistic, as well as a set of chosen multimodal features. An overview of our feature set is depicted in Figure 1. Since we are dealing with educational videos, we assume that for each data sample a video file is available, together with a PDF file of the shown presentation as well as a speech transcript.

### 2.1 Unimodal Features

There are three kinds of unimodal features: audio, linguistic, and visual.

**Audio Features.** The openSMILE-toolkit [11] is used to extract the audio features, except for the pitch variation information which is extracted according to Hincks [16]. We selected the feature subset of the ComParE challenge (6,373 dimensions, around 70 low-level descriptors with multiple features each) and reduced it to nine features and their arithmetic mean, see Table 1. For our study, we selected those features that have either shown an impact on the audio quality before (Jitter, F0 Harmonics ratio, ...) or are very likely to influence the perceived quality of the audio (Energy, Loudness, Harmonics-to-Noise Ratio, etc.).

| Feature                              | Description                                                |
|--------------------------------------|------------------------------------------------------------|
| Loudness                             | Sum of a simplified auditory spectrum                      |
| Modulated loudness                   | Sum of a simplified RASTA-filtered auditory spectrum       |
| Root-Mean-Square energy              | Square root of mean of the discrete values of the sound pressure |
| Jitter                               | Deviation from true periodicity of a presumably periodic signal |
| ∆Jitter                              | Normalized average length deviation from true periodicity of a presumably periodic signal |
| Shimmer                              | Amplitude variation of consecutive voice signal periods     |
| Harmonicity (spectral)               | Ratio between the minima and the maxima in relation to the amplitude of the maxima from a magnitude spectrum |
| Logarithmic Harmonics-to-Noise Ratio | Logarithmic scale of the ratio harmonic to noise component in the wave signal |
| Pitch Variant Quotient              | Standard deviation of the pitch, which is divided by the mean of the pitch (cf. Hincks [16]) |

Table 1: Automatically extracted audio features.
Linguistic Features. The extraction of linguistic features aims to describe how the presenters articulate themselves by means of syllable duration and speaking rate. De Jong and Wempe’s [9] Praat [4] script was used to extract these features, namely speech rate, articulation rate, and average syllable duration (ASD). All of them are derived from the number of vowels or syllables per time interval and indicate if the speaker is talking too fast or too slow.

The video transcript does also contain a lot of useful information regarding speech quality. Since the influence of textual content on the knowledge gain has been examined extensively (e. g., by Gadiraju et al. [12]), this information is disregarded here and the focus set on the lesser researched features. However, we make use of the content of the transcript in a different way in section 2.2.

Visual Features. For the visual content, we examine the PDF files of the presentation slides. With the bash command pdftotext we extract text layout information from the PDF slides. The command extracts the position of each text element as well as the size of the slide. The elements of a slide are stored in an hierarchical way starting with the biggest text element, which contains multiple text lines and each line consists of multiple words. This information is converted into a XHTML file. Similarly, the pdftohtml command is used to extract the image position and size of the slide, which are stored in an XML file. The generated files are then parsed to JSON, since the format is more convenient for data handling.

Based on this representation, we compute two features related to the design of the slides, which are text ratio and image ratio. They describe, how much slide space is covered by each of the modalities according to Formula 1.

\[
TextRatio = \frac{\sum_{i=1}^{n} TextArea_i}{Area_{slide}}.
\]

Also, for each file we store the mean and sample variance of the text ratio and image ratio values of all slides.

2.2 Cross-modal Features

In this section, we present a set of multimodal features which aim to model specific quality aspects of a presentation. We were inspired by criteria that are important to us humans, for instance, the way and frequency the presenter highlights important aspects on the slides. If we are able to capture these metrics, we can rank videos with similar content according to their presentation quality providing an optimal recommendation to a learner.

Highlight of Important Statements. This feature is supposed to indicate how often important statements are emphasized per slide and over the complete slide set. To identify the text boxes most likely containing the key components of a slide, we use the information from the document layout analysis, which we stored in JSON format earlier.

In this procedure, we use the following natural language processing functions which were adapted from Bird et al. [3]:

- N(): return a list of nouns from a sentence
- LEM(): return a set of lemmas from a list of words
- SYN(): return a set of synonyms from a list of words

Our assumption for the identification of important text is that font size is often proportional to importance. Since we do not have the font size information for each slide we sort the text lines by the area they cover. However, simply choosing the area biggest text areas does not yield good results, because there are often bullet points of similar importance but cover a text areas of different size. So we cluster the text areas according to the following rule: For each text area starting from the biggest one, if the area difference to the next biggest text area is smaller than &% of the slide size (in our experiments &%), add it to the existing cluster, otherwise create a new one. All text areas in the two biggest clusters are considered to contain important statements. This usually includes the text area of the title and all headlines of the highest category. For each selected text line, we first extract the sentence(s) (Stimp) from the respective JSON file. Then, the nouns and their synonyms are extracted from the text. Finally, we lemmatize the nouns and their synonyms:

\[
St = LEM(N(Stimp) \cup SYN(N(Stimp))).
\]

Locating Emphasized Transcriptions. This feature is designed to capture the ability of the presenter to consider important statements shown on the slides as well as their emphasis through his or her voice. If so, there should be a corresponding local maximum in the audio signal. To get this information, we need to align the speech transcript with the audio signal in the time frame where the currently observed slide was covered. The segmentation of the video according to single slides is done manually. For each slide and associated time frame we need to find the corresponding segment of the speech transcript. The transcript is segmented into blocks of 10 seconds, which is a common interval in speech analysis (cf. Hincks [16]). A slide segment of arbitrary length can contain multiple of these blocks and it is important to find the correct one for each highlighted statement, see Figure 2. This is done using the audio signal. We encoded the audio information via three metrics:

\[
F_0, loudness, and energy. If all of them have a maximum at around the same timestamp, we assume the presenter emphasized the word said at that moment. Locating the exact words at that moment is done by choosing the speech transcript block whose temporal centre is closest to the found maximum. The list of these presumed important statements are stored in EmphasizedTranscriptions and we lemmatize them again:

\[
Tr = LEM(EmphasizedTranscriptions).
\]

Finally, the fraction of highlighted statements is calculated as:

\[
Highlight = \frac{|St \cap Tr|}{|St|}.
\]
**Level of Detailing.** Another possible measure of quality for the presentation is the overlap of spoken text with the information on the slide. As literature from video learning suggests [5], audio and visual contents should provide complementary information rather than being overly redundant. Thus, we examine if the speaker only reads the information already present on the slide, or if the oral explanation provides further detail, giving an appropriate amount of additional information. For this purpose, we calculate the ratio of said words to shown words on the slide.

Again, we use the speech transcript blocks from the previous metric to count the number of words said during the time frame when the corresponding slide was visible. Every speech transcript block overlapping with the duration of the slide is considered. Blocks at the interval boundaries are cut off appropriately. If a transcript block overlapped 70% at the end of a slide and contained 10 words, we would consider the first seven words and would dismiss the last three. All found words are therefore stored in $\textit{Words}_{said}$. The number of words on the slide $\textit{Words}_{slide}$ were gathered from the process explained in section 2.1.

The **level of detailing** is calculated as the ratio of number of said words to number of words on the slide, see Formula 5:

$$\text{Level Of Detailing} = \frac{|\textit{Words}_{said}|}{|\textit{Words}_{slide}|}.$$  

(5)

Also, for each video the mean and sample variance of these values was calculated for later usage.

**Coverage of Slide Contents.** This metric encapsulates if the speaker talked about all the information shown on the slide of if some parts were skipped. Again, this gives an idea if the overall talk is well structured and timed or rushed and hard to follow for an observer, since some information shown is left without explanation. We can reuse the $\textit{Words}_{slide}$ from the previous section. For the words said by the speaker, we reuse the already established $\textit{Words}_{said}$ from the previous metric.

The words in $\textit{Words}_{slide}$ and $\textit{Words}_{said}$ were lemmatized again, which enables an easier comparison. The **Coverage** of slide content is calculated as the ratio of the number of common words in $\textit{Words}_{said}$ and $\textit{Words}_{slide}$ to the total number of words on the slide:

$$\text{Coverage} = \frac{|\textit{Words}_{said} \cap \textit{Words}_{slide}|}{|\textit{Words}_{slide}|}.$$  

(6)

Similarly, the mean and sample variance of the values of all slides are calculated for later usage. The full list of manual and automatic features can be seen in Table 2.

### 3 DESIGN OF THE USER STUDY

To estimate the expressiveness of our automatically extracted features we conducted a user study. We aimed to get human ratings for different quality aspects of lecture videos, while these aspects are covered by respective feature sets. In addition, we asked for an overall rating of a lecture video. Furthermore, every participant was asked to fill in a knowledge test before (pre-test) and after (post-test) watching the video, aiming to measure the capability of a video to convey knowledge. We conducted a correlation analysis to find out which features are correlated with the quality aspects of lecture videos as well as with knowledge gain.

#### 3.1 Data Acquisition

Our dataset consists of 22 videos (with associated slides and speech transcripts) from edX1. The course materials of this course are Copyright Delft University of Technology and are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License [8]. The available course materials on edX are provided the following formats: videos in MP4, slides in PDF, and transcriptions in SRT. We chose this source since it does not require any further pre-processing, it is open access and the speech transcript is of high quality (presumably they have been manually reviewed).

#### 3.2 Participants and Task

The subject of the 22 videos is software engineering. Each video has exactly one presenter, while the full dataset has nine different presenters with varying slide designs. We employed 13 participants (10 men, 3 women) from our university with a computer science background, an average age of $25.8 \pm 2.4$ years, and asked everyone to watch and assess nine videos. With an average video length of 8 minutes and the time to fill out the evaluation forms the full experiment took 1.5 – 2 hours. The participants were rewarded 13 Euro/Hour. We made sure that each set of videos contained as many different presenters as possible and that each video is viewed at least by five different people. We chose to gather multiple ratings for the same video instead of investigating a larger set of videos to be more robust against outlier ratings.

---

1[https://courses.edx.org/courses/course-v1:DelftX+GSE101x+1T2018/course/](https://courses.edx.org/courses/course-v1:DelftX+GSE101x+1T2018/course/)
3.3 Experimental Setting

A common way to estimate the knowledge gain of a participant during a learning session is to conduct a knowledge test before and after a controlled learning session (e.g., Yu et al. [27]). The resulting score difference indicates how much was learned. Even though the potential knowledge gain depends heavily on the participant, by choosing a subject that is most likely unfamiliar to a majority of people, we try to circumvent this problem. It is however important to ensure that the participants have a chance to understand the content, otherwise the knowledge gain will be low again. Therefore, we selected the topic *Globally Distributed Software Engineering*. It is on one hand, a computer science topic related to the studies of our participants but also a very specific area which is not part of their curriculum. Thus, everyone had a chance to understand the topic based on their prior knowledge and therefore favouring a positive knowledge gain.

A negative effect of the pre-test is that it might influence the user behavior by providing hints on what to focus on in a video, because participants will try to get a good score on the post-test. We gathered a set of relevant questions inspired by the intermediate quizzes in the course material. However, we made sure to amend and change them since their reuse is restricted. We chose two to four questions intended to be asked after each video with a similar amount of unrelated questions from other videos. Also, we put the videos in random order so it was hard to guess which of the questions will be the relevant ones. In addition, the number of possible answers was different every time. Exemplary, the knowledge test for video 6_2a can be seen in Figure 3.

After filling out the pre-test, the participant was instructed to watch the entire video without pausing, rewinding, or taking notes. Reason for that is that we wanted the participants to get a full impression of the presentation instead of, again, just skipping to the relevant parts for the knowledge test to get a good score. Similarly, we assumed that when we allowed people to take notes, they would just write reminders down about the pre-test and focus solely on their appearance in the video. Admittedly, this is slightly different to a realistic setting, but we applied it in favor of the knowledge gain measurement. After watching the video, the person is asked to answer the same questions again and also to fill out an evaluation form with questions that are related to different quality aspects, see Table 3. The items are assessed using a Likert scale from 1-5.

This paragraph describes how we scored the knowledge test with the goal of treating each video with a similar importance, independent of 1) the number of relevant questions per quiz and 2) the number of possible answers per question. First, the score for an unanswered question will be treated as zero since we gave the participants the option to skip a question in order to discourage random guessing. If the question was answered, we calculate the score for each answer option by increasing (decreasing) the score by 1 for a correct (false) answer. Thus, a question with five answer options can yield the following scores: −5, −3, −1, 0, 1, 3, 5.

The knowledge gain of participant \( s \) after watching video \( v \) is then calculated as the difference between the pre-test score \( PB_{vs} \) and post-test score \( PA_{vs} \). Let \( n_v \) be the number of participants who watched video \( v \).

We start by computing the mean knowledge gain of participants for each video:

\[
\mu = \frac{\sum_{j=1}^{n_v} (PA_{vs} - PB_{vs})}{2n_v}.
\]  

Next is the standard deviation of the knowledge gain:

\[
\sigma = \sqrt{\frac{1}{2n_v} \left[ \sum_{x=1}^{n_v} (PB_{vs} - \mu)^2 + \sum_{x=1}^{n_v} (PA_{vs} - \mu)^2 \right]}.
\]  

Based on the mean and standard deviation value, the scores are normalized. \( PB'_{vs} \) is the normalized score which is computed by subtracting the mean and dividing by the standard deviation. The same applies for \( PA'_{vs} \):

\[
PB'_{vs} = \frac{PB_{vs} - \mu}{\sigma}.
\]
Consequently, the normalized knowledge gain of participant $s$ for video $v$ is:

$$KG_{Us} = PA_{Us} - PB_{Us}. \quad (10)$$

$KG_{0}$, the overall knowledge gain for video $v$ is finally calculated by the average of all participants’ knowledge gain:

$$KG_v = \frac{1}{N_v} \sum_{s=1}^{N_v} KG_{Us}. \quad (11)$$
While the latter ones lack an intuitive explanation, the features associated with a more lively voice. This inconsistency may be due to the data smoothing for \( F_0 \) values. The smoothed values are less accurate and could lead to this effect.

**Vocal Diversity.** The Vocal Diversity of the speaker had a strong correlation with three audio features and two multimodal features. While the latter ones lack an intuitive explanation, the features from the audio signal, namely logarithmic HNR, \( F_0 \), and speech rate, most certainly directly contributed to this measurement. For instance, Yumoto et al. [28] found a negative correlation between Harmonics-to-Noise Ratio and hoarseness, hinting that this metric resembles the voice quality of the speaker. For the negative correlations, shimmer, as mentioned before, severely impacts this metric and should definitely be considered when trying to automatically predict it.

**Speed of Presentation.** The rating regarding an appropriate presentation speed positively correlated with the two jitter features, which indicates changes in the speech rate. Possibly, the presenter regularly adapted the presentation speed based on the difficulty of the current slide – which would be an indicator for a good presentation. It is negatively correlated to the amount of images on the slides, which can be explained by the fact that an excessive amount of images on slides adds to the cognitive load: without sufficient time to process visual contents, the viewer might feel overwhelmed.

**Filler Words.** Our filler words quality metric has a strong correlation with the articulation rate feature. This can be explained by the fact that a high articulation rate represents a high amount of syllables per time unit, which is especially apparent if the speaker uses a lot of filler words like “uhhhmm” or “ehhmm”. The remaining features associated with it do not allow for a direct interpretation but might be correlated indirectly.

**Appropriate Detailing.** Our cross-modal feature appropriate detailing has a strong correlation with the variance of its automatic counterpart, namely detailing variance (\( \alpha < 0.05 \)). This is interesting since it indicates that the participants valued a high variance in the detailed parts of the speech more than a presentation where it happened constantly. In other words, at least for our dataset, a good presenter should only detail the necessary sections of the content.

### 4.3 Correlation analysis for features and knowledge gain

One of the most important features for an educational video is its ability to convey the intended information and, consequently, to improve the viewers’ knowledge state. The following section highlights the automatically extracted features that had the biggest impact on the knowledge gain of the participants during the user study. Since both our extracted features and the knowledge gain scores are interval data, Pearson correlation analysis is used.

First, Figure 5 shows the distribution of achieved knowledge gains for all 23 relevant questions answered during the user study. The data is normalised preventing questions with more possible answers to skew the results. The average knowledge gain of +0.48 shows that the chosen topic enabled all participants to improve their knowledge in this area to a certain degree. Also, the questions were neither to easy nor to hard, which would have also resulted in a low average knowledge gain. However, the individual knowledge gain of the participants varied significantly (avg. standard deviation is 0.36), reflecting the subjectivity of the learning process.

Table 6 shows the correlation between features and knowledge gain and the corresponding confidence level \( \alpha \). The feature with the biggest positive impact on the knowledge gain with a value of \( r_p = -0.357, \alpha < 0.2 \) is Modulated Loudness. This reflects the intuitive assumption that loud voices or unpleasant background noises hinder learning experience and effect. Knowledge gain also correlates negatively with Image Ratio Variance (\( r_p = -0.282, \alpha < 0.3 \)),

| Human-rated Aspect | Automatic Features | \( r_s \) | \( \alpha < \) |
|--------------------|--------------------|----------|---------|
| **Clear Language** | RMS Energy         | 0.368    | 0.1     |
|                    | Loudness           | 0.430    | 0.05    |
|                    | Harmonicity (spectral) | 0.435 | 0.05    |
|                    | Detailing Mean     | 0.541    | 0.02    |
|                    | Detailing Variance | 0.448    | 0.05    |
|                    | \( F_0 \)           | 0.608    | 0.005   |
| **Vocal Diversity**| Log. HNR            | 0.389    | 0.01    |
|                    | \( F_0 \)           | 0.429    | 0.05    |
|                    | Detailing Variance | 0.463    | 0.05    |
|                    | Speech Rate        | 0.525    | 0.02    |
|                    | Detailing Mean     | 0.615    | 0.005   |
| **Summary**        | \( \Delta \) Jitter | 0.408    | 0.1     |
|                    | Jitter             | 0.423    | 0.1     |
| **Filler Words**   | Image Ratio Mean   | 0.394    | 0.1     |
|                    | Articulation Rate  | 0.459    | 0.05    |
|                    | Detailing Variance | 0.493    | 0.05    |
|                    | Detailing Mean     | 0.601    | 0.005   |
| **Image Design**   | ASD                | 0.371    | 0.1     |
| **Appropriate Detailing** | Detailing Variance | 0.378    | 0.1     |

Table 4: Positive correlation results with \( \alpha < 0.1 \) (\( r_s \): Spearman Correlation Coefficient, HNR: Harmonics-to-Noise Ratio, Detailing: Level of Detailing)

| Human-rated Aspect | Automatic Features | \( r_s \) | \( \alpha < \) |
|--------------------|--------------------|----------|---------|
| **Clear Language** | PVQ Average        | -0.416   | 0.1     |
|                    | Coverage Variance  | -0.489   | 0.05    |
|                    | shimmer            | -0.623   | 0.005   |
| **Vocal Diversity**| Text Ratio Mean    | -0.396   | 0.1     |
|                    | Highlight          | -0.406   | 0.1     |
|                    | shimmer            | -0.437   | 0.05    |
| **Overall Rating** | Speech Rate        | -0.455   | 0.05    |
| **Image Design**   | Articulation Rate  | -0.368   | 0.1     |
| **Speed of Presentation** | Text Ratio Variance | -0.454 | 0.05    |
|                    | Image Ratio Variance | -0.561 | 0.01    |
| **Text Design**    | Image Ratio Mean   | -0.466   | 0.05    |
| **Structure**      | Coverage Variance  | -0.402   | 0.1     |
| **Summary**        | Coverage Variance  | -0.457   | 0.05    |

Table 5: Negative correlation results with \( \alpha < 0.1 \) (\( r_s \): Spearman Correlation Coefficient, Coverage: Coverage of Slide Contents, Structure: Structuring of Presentation)
5 CONCLUSIONS

In this paper, we have presented a set of unimodal and cross-modal features that can be automatically extracted from lecture videos. Further, we have presented a user study that investigated the correlations between these features and quality aspects of the lecture video. Also, the knowledge gain of users was measured and the correlations to the video features were evaluated. The results provided insights about a number of moderate to good correlations.

We were able to represent the objective quality metrics Clear Language, Vocal Diversity, Speed Presentation and Filler Words each with at least three automatically extracted features and a confidence level of at least 95%. Additionally, we presented an approach to evaluate the level of Appropriate Detailing which correlated with human assessment.

In the future, we will investigate if and how these results generalize to other areas.

Also, it is possible to measure the impact of user-related features like click behavior or gaze tracking data on the learning process and their correlation to our quality metrics.

Finally, we plan to utilize machine learning methods in order to predict video quality based on our set of features. This information could serve as an additional input for a recommender or retrieval system and add another dimension to distinguish between videos of similar content but differing presentation quality. In this way, exploratory search capabilities of educational video portals and MOOCs could be improved.

ACKNOWLEDGMENTS

Part of this work is financially supported by the Leibniz Association, Germany (Leibniz Competition 2018, funding line “Collaborative Excellence”, project SALIENT [K68/2017]).
