Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Verteilte Systeme
Fachgebiet Visual Analytics

Automatic Quality Assessment of Lecture Videos Using Multimodal Features

Masterarbeit

im Studiengang Informatik

von

Jianwei Shi

Prüfer: Prof. Ralph Ewerth
Zweitprüfer: Prof. Joel Greenyer
Betreuer: M.Sc. Christian Otto

Hannover, 23.05.2019

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 23.05.2019

Jianwei Shi

Zusammenfassung

Multimedia Retrieval, eine entwickelte Methodologie, welche aus Information Retrieval stammt, wird in der digitalisierten Gesellschaft weit verbreitet eingesetzt. Bei der Suche nach Videos im Internet, müssen diese nach ihrer Relevanz sortiert werden. Die meisten Ansätze berechnen die Relevanz jedoch nur aus grundlegenden Inhaltsinformationen. Ziel dieser Arbeit ist es, Relevanz in verschiedenen Modalitäten zu analysieren. Für den konkreten Fall von Vortragsvideos, Merkmale von folgenden Modalitäten werden von dementsprechenden Kursmaterialien extrahiert: akustische, linguistische, und visuelle Modalität. Außerdem sind modalitätsübergreifende Merkmale insbesondere in dieser Arbeit zunächst vorgeschlagen und berechnet durch die Verarbeitung von Audio, Bilder, Transkripte und Texte. Eine Benutzerevaluation wurde durchgeführt, um Benutzermeinungen in Bezug auf die erzeugten Merkmale zu erheben. Die Ergebnisse haben gezeigt, dass die meisten Merkmale ein Video in verschiedenen Aspekten widerspiegeln können. Die Art und Weise, wie der Lerneffekt durch diese Merkmale beeinflusst wird, wird ebenfalls berücksichtigt. Für die weitere Forschung baut diese Studie eine solide Basis für die Extraktion der Merkmale auf. Zudem gewinnt die Arbeit ein besseres Verständnis zum Lernen.

Abstract

Mutimedia retrieval, a developed methodology based on information retrieval, is broadly used in the digitalised society. When searching videos online, they need to be sorted according to their relevance. However, most approaches calculate the relevance only from basic content information. This thesis aims to analyse the relevance in multiple modalities. For the specific case of lecture videos, features from following modalities are extracted from corresponding course materials: audio, linguistic, and visual modality. Furthermore, cross-modal features are specifically first proposed in this thesis and calculated by processing audio, images, transcripts, and texts. A user evaluation has been conducted to collect user's opinions with regards to these generated features. The results have shown that most features can reflect a video in multiple aspects. The way the learning effect is influenced by these features is considered as well. For further research, this study builds a solid base for feature extraction and gains a better understanding of learning.

Contents

Chapter 1

Introduction

Nowadays, information retrieval applications are omnipresent in the world of digitalisation. Useful information can be found by simply entering a query in a web search engine. Search results are usually ranked by their relevance, so that the most useful information is shown at the beginning. In this way, users can quickly find the information they need. At the same time, corresponding advertisements can be recommended in the result list. Advertisement providers want to reach their target group accurately, so the advertisements should be closely related to the query. All these processes are based on analysis of correctness for the results, which is a fundamental part of a good search engine.

In classic search machines, the organisation of documents and their ranking work as follows: pages (documents) in internet are crawled and downloaded at first; a document is then divided into terms; each term is evaluated based on frequency, document structure, and special properties such as HTML tag; after that, the corresponding documents are retrieved; documents in result set are ranked according to evaluations of contained terms; for link-based ranking, the hypertext links between documents are used to calculate the relevance. (cf. Bauer [2])

Terms and hypertext links are extracted and analysed in the above mentioned classic search engines. Terms consist of words, symbols, and other characters; links are WWW addresses: all of these are in fact textual information.

Consider this: How does the ranking work if videos are searched? Let us have a look at two commercial video hosting platforms. The videos at Youtube[1] can be sorted by relevance, upload date, number of calls, and duration. Similarly, the criteria at vimeo[2] are relevance, recently uploaded, popularity, title (A - Z), title (Z - A), longest, and shortest. The criteria, except relevance, are all about representations of a video, i.e., metadata.

[1] `https://youtube.com`, accessed 5-May-2019
[2] `https://vimeo.com`, accessed 5-May-2019

Criteria about video content itself are missing.

A video represents information in textual, aural modalities, and other aspects. There is not only text, but also voice of speakers, images etc. in a video. A human being perceives all modalities of a video and will have an overall judgement on it. The machine which does the video retrieval should do exactly the same: A video should be analysed in multiple aspects. Explicit multimodal criteria, such as the engagement of speaker and overall rating, should be considered.

The interdisciplinary research project "Search as Learning – Investigating, Enhancing and Predicting Learning during Multimodal Web Search" [14] aims to explore how learning processes on the web work and how they can be improved. A challenging part of the project is the recommendation system. For one user with specific learning tasks, the system should recommend the videos which best fit the requirement.

My work facilitates this research by exploring how videos can be automatically assessed from multiple aspects. The assessment serves as a dedicated criterion for ranking of videos, so that users are supposed to have greatest knowledge gain from seeing the first few videos in the recommendation list.

This thesis is structured as follows: Chapter 2 will have an overview of related work. Chapter 3 begins with an general introduction of all features across different modalities. After that, the feature extraction for each modality will be explained. The remaining part of the chapter deals with cross-modal features. Chapter 4 focuses on the organisation of a user study. Chapter 5 discusses the correlation between machine generated features and human evaluation. The last chapter concludes the paper and gives an outlook.

Chapter 2

Related Work

Before explaining how multimodal automatic quality assessment works, it is necessary to have a look at existing work for this field.

Multimodal metadata analysis is a popular research topic. This chapter begins with applications for multimodal assessment. Based on the multimodal assessment, there are many mature multimodal content-based video retrieval approaches. After that, this chapter dives into specific modalities for low-level features. At first, feature extraction for different modalities will be discussed. Having all features from different modalities, the related work of further feature organisation is then introduced. The last part shows works about the evaluation of presentation videos.

2.1 Multimodal Embeddings and Their Applications

The sentiment analysis by Poria et al. [22] was conducted with multimodal features. The videos were analysed in aspect to audio, visual, and textual modalities. The proposed system was outperforming all state-of-the art systems by more than 20% with regards to accuracy.

Going to presentation field, Haider et al. [15] focused on prosodic and visual features from presentations. The presentation quality was predicted with high accuracy.

In the work from Balasubramanian et al. [1], so-called "multimodal metadata" is extracted. Similarily, the metadata is extracted from audio transcripts and slide content. They claim that their multimodal approach helps user to use the retrieval system with more convenience.

In the field of content-based video retrieval, the input of the system comes from multimodal sources. The content-based video retrieval relies on text information from multiple sources.

Looking to find a more efficient method for video retrieval on the internet, Yang and Meinel [28] propose a content-based video retrieval method. Not

only visual but also audio resources are used as an input. The metadata, key words or others based on the content from lecture videos, are extracted. On the one hand, the text, slide structure etc. can be extracted from visual resources. On the other hand, transcription is generated from the audio signal. It is proven in their evaluation that the suggested retrieval method can enrich the understanding level for the contents.

2.2 Video Assessment

This section gives an overview of video assessment in different aspects. Firstly, single modal features are introduced. Secondly, features are further organised to find relationships between them. Lastly, evaluation criteria and approaches of automatic evaluation are concerned.

Single Modality

Eyben et al. [12] have launched an open-source toolkit for emotion and affect recognition, with which the basic audio features can be extracted. In his dissertation [10], the upgraded toolkit openSMILE has been introduced. At the same time, the calculation of Low-Level Descriptors (features) from audio have been explained in detail.

For the speech rate calculation, de Jong and Wempe [8] have written a script in the software program Praat [4]. Without generating transcriptions, the speech rate feature was extracted directly from audio.

Combinations of Features

Haider et al. [15] have conducted an investigation of prosodic features. The prosodic features are complete feature set of ComParE challenge [24], perceived loudness, and vocalisation to pause ratio. In addition, visual features from hand gestures are extracted. They have used the correlation matrix to find the relation between prosodic and visual features.

Similarly, Chen et al. [6] have used multimodal sensing to give presentation quality. The speech features, body movement features, and visual features from presentation slides were extracted. They have used the Principle Component Analysis on human scores to address the two main modalities of presentation videos. Component 1 is for delivery skills, which means the voice information, body language and so on. Component 2 is for slides quality, with regards to grammar, readability, and visual design. For each component, Pearson correlation was used to get the relation between different features.

Evaluation for Lecture Videos

The Massive Open Online Course (MOOC) platform serves for learning by offering lecture videos and other course materials. Yousef et al. [29] have carried out an empirical study of the quality assessment for MOOC platforms. Different aspects are considered, including video content category.

Several works provide scoring model for assessment of lecture videos, which is generated using machine learning algorithms.

Li et al. [20] have designed a set of assessment rubrics to evaluate a lecture video. For multimodal features, a multi-stream deep learning framework was developed. The features came from video and skeleton modality. The video frames were cropped based on skeleton features. A convolutional neural network (CNN) was used to extract features from every video frame. The features from every video frame was weighted with a temporal attention module. Then the features were fed to a framework of long short-term memory (LSTMs). The evaluation has shown that (a) the deep learning approach gave better results than the previous method [13]; (b) for the video modality, proposed attention mechanism had significant improvement in comparison to no attention mechanism.

In order to predict human scores from presentations, Chen et al. [6] used support vector machine (SVM) regression and stochastic gradient boosting (GBM) to train the scoring model. According to their experiment, GBM models showed higher performance than SVM models.

Taken together, these studies support the notion that multimodal features are an important data source for evaluation. Features with correlations can be further organised for training and assessment. The trend using machine learning algorithms can help to generate a scoring model which can give evaluation automatically.

In the next chapter, I will present my own procedures and methods based on these related works.

Chapter 3

Feature Extraction

In order to assess videos, it is necessary to find descriptors, which can be recognised by computer. These descriptors are called features. This chapter deals with multimodal feature extraction from videos. Firstly, a general perspective for all multimodal features is given. At the same time, the organisation of video dataset is introduced. The second section is about extraction logic of single modal features. Finally, cross-modal features are discussed.

3.1 Overview of All Features

Before introducing all features, it is important to have a look at data preparation.

3.1.1 Data Preparation

The lecture videos were selected based on their usefulness for analysis. In order to make a dataset for lecture videos, the course materials from edX[1] are used. The available course materials on edX are: videos in MP4 format, slides in PDF format, and transcriptions in SRT format.

In a video, a presenter appears in it and shows the corresponding slides. The transcription precisely indicates what a presenter has said during a certain timeframe. So the videos can be analysed according to a variety of modalities.

The subject of the selected course is software engineering. A complete list of courses is attached in appendix A.

[1] `https://www.edx.org`

3.1.2 All Features in General

Good quality videos should have the following characteristics (among others):

1. voice of speaker can be understood well;
2. speaker has used a variable and appropriate sound, instead of a monotonous one;
3. speed of the speaker is appropriate;
4. speaker has spoken fluently with few filler words;
5. images or slides in the video are easy to follow;
6. important statements are emphasised;
7. slide contents are all covered;
8. lecture content is explained in appropriate detail.

I have adapted three characteristics (1, 6, and 8) based on the study from Yousef et al. [29]. Additionally, I have extended the list by other points.

In order to give feedback for the above eight points, corresponding features from different modalities need to be extracted and analysed. All related features are listed in figure 3.1.

Four different features are related to these eight points: audio (1 and 2), linguistic (3 and 4), visual (5), and cross-modal (6, 7, and 8) features. Audio features is about what the audience has heard, e.g., how powerful the sound is; linguistic features describe what the used language looks like, e.g., how fast a presenter speaks; visual modality is related to the what the audience has seen, e.g., how much text on the slide is. Based on these features from multiple modalities, the cross-modal features are considered as well, which are explained in the last section.

In the next section, single modality features are explained in detail.

3.2 Single Modal Features

This section introduces single modal features which are generated from audio data. To make feature extraction possible, the WAV files were converted from MP4 data.

3.2.1 Audio Features

Following audio features are extracted using toolkit openSMILE from Eyben et al. [11]: fundamental frequency (F_0), loudness, modulated loudness, Root

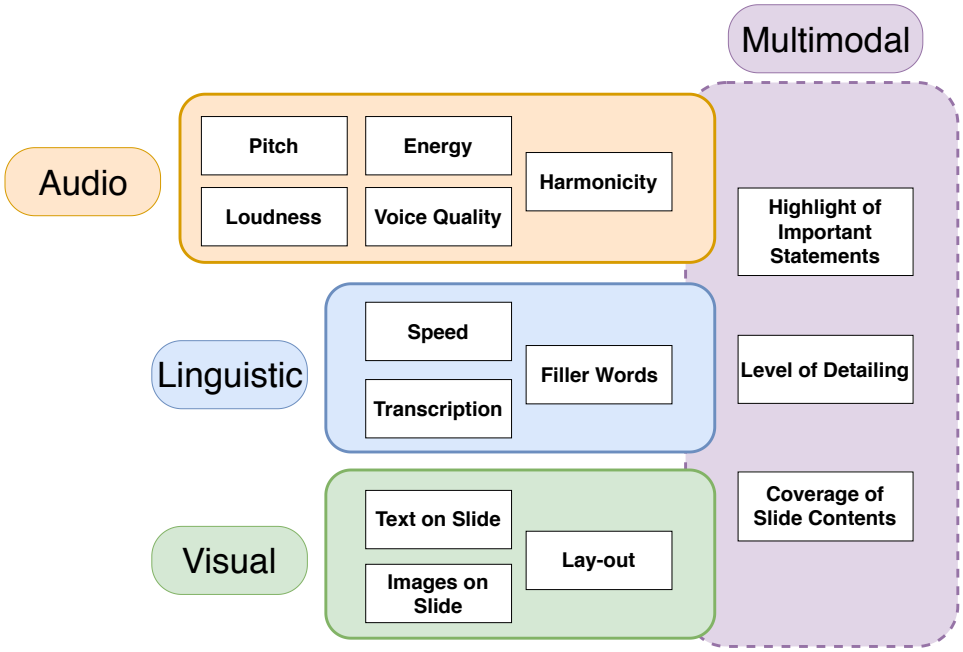

Figure 3.1: A Bird's Eye View of Features

Mean Square (RMS) energy, jitter, δ jitter, shimmer, harmonicity (spectral) and logarithmic Harmonics-to-Noise Ratio.

These features are part of the ComParE challenge feature set from Schuller et al. [24]. For a video, values of these features are extracted per one second and later saved in JSON files. At last, the average values for each audio feature are calculated.

Specifically, for calculating the Pitch Variant Quotient average, F_0 is extracted again every 0.01 second.

The definitions of each feature are explained in the following subsections.

Fundamental Frequency

The F_0 is defined as the lowest frequency of a periodic waveform, which describes a voiced speech in a basic way. The calculation with openSMILE works as follows: The spectral signal is transformed at first; and then the pitch is detected using Subharmonic Summation described from Hermes [17]; the pitch contour will then be smoothed; lastly, the final frequency value is extracted. For unvoiced regions, the value of F_0 is zero.

Based on F_0 feature, the Pitch Variant Quotient average is calculated.

Pitch Variant Quotient Average

In order to represent the pitch variation information, Hincks [18] has proposed the Pitch Variant Quotient (PVQ) which is based on statistical analysis.

In her study, the pitch value (F_0) is extracted for every 0.01 s from the audio signal. After that, all values of zero are deleted. The error values, which are evidenced by visual inspection, are also deleted. For each ten seconds of speech, the standard deviation (sd) and the mean are calculated from F_0 value list. The PVQ for a part of speech is defined as:

$$PVQ = \frac{sd}{mean}. \tag{3.1}$$

Ten seconds is selected as an ideal interval for analyse "because it was enough time to guarantee the inclusion of a fair amount of speech at normal pausing rates" (cf. Hincks [18]).

As the smoothing is used for F_0 extraction, the visual inspection is omitted in my study. The algorithm 1 has shown the steps. After calculating PVQ values for all parts of speech, the **PVQ average** of the values are calculated as the feature for whole video.

Algorithm 1: Calculation of Pitch Variant Quotient Average

input : complete F_0 value list `f0List`
output: PVQ average

```
pvqList ← empty list;
delete all elements with zero value from f0List;
divide f0List into 1000-element sublists sequentially;
// except the last sublist which contains no more than 1000 elements
f0SubListGroup ← sequentially ordered sublists;
foreach sublist sl in f0SubListGroup do
    sd ← standard deviation of values in sl;
    m ← mean of values in sl;
    pvq ← sd/m;
    append pvq to pvqList;
end
return mean of values in pvqList;
```

Loudness

The loudness indicates sound power of the human perception. There are two features related to loudness, which are called **loudness** and **modulated loudness**.

These two features are described as follows (cf. Eyben [10], pp. 260-261):

> The loudness is computed as the sum of a simplified auditory spectrum [...]
>
> The modulated loudness is computed as the sum of a simplified RASTA [RelAtive Spectral TrAnsform, see Hermansky et al. [16]] filtered auditory spectrum [...] This loudness measure reflects the loudness contained in the speech signal while suppressing loudness influence from near stationary or high-frequency noise.

Energy

Signal energy is a very simple audio descriptor which indicated the strength of the speaking voice. The **RMS energy** is chosen as measurement.

For signal x, $x(n)$ is the amplitude for a sample n, the RMS Energy [19] for samples $n \in [1, N] \cap \mathbb{Z}$ is defined as:

$$E_{rms} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x^2(n)}. \tag{3.2}$$

Jitter and Shimmer

The jitter- and shimmer-related characteristics can describe the extent of sickness of a voice.

For diagnosis of pathologic voice, the features which relate to jitter and shimmer are used in a study from Teixeira et al. [26]:

> Jitter is defined as the parameter of frequency variation from cycle to cycle, and shimmer relates to the amplitude variation of the sound wave, [...] The jitter is affected mainly by the lack of control of vibration of the cords; [...] The shimmer [...] is correlated with the presence of noise emission and breathiness.

Three related features are extracted: **absolute period to period jitter**, **δ jitter**, and **absolute period to period shimmer** (cf. Eyben [10]).

Harmonicity

Harmonicity indicates the signal quality. There are two features describing that: **harmonicity (spectral)** and **logarithmic Harmonics-to-Noise Ratio** (log. HNR).

The calculation of Harmonicity (spectral) is explained as follows (cf. Eyben [10], p. 43):

> Harmonicty is computed directly from a magnitude spectrum by applying a simple peak picking algorithm based on identification of local minima and maxima [...] Then, the ratio between the minima and the maxima in relation to the amplitude of the maxima is computed.

Harmonics-to-Noise Ratio (HNR) is defined as the ratio harmonic to noise component in the wave signal. The logarithmic scale of it (log. HNR) is calculated for a better representation. (cf. Eyben [10], pp. 77-78)

3.2.2 Linguistic Features

These features are extracted from the language level: speech rate, articulation rate, and average syllable duration. Transcription of videos is also an important linguistic feature, which is represented in a SRT file. Subtitles and their corresponding timestamps indicating begin and end are stored in the file. Transcription information will be later used in cross-modal feature extraction. The following part concentrates on the extraction of the first three features.

Speech-related Features

A Praat [4] script adapted from de Jong and Wempe [9] is used to generate speech related features. All of them are based on information about vowels within a syllable, which are called syllable nucleus.

At first, the syllable nucleus is recognised. Having the whole duration $TotalTime$ of the speech, the **speech rate** is calculated as the number of syllable nuclei in a unit time:

$$SpeechRate = \frac{|SyllableNuclei|}{TotalTime}. \tag{3.3}$$

As there can be pauses during the speech, the time when the speaker is actually speaking is calculated as $PhonationTime$. The number of syllable nuclei in unit $PhonationTime$ is called **articulation rate**:

$$ArticulationRate = \frac{|SyllableNuclei|}{PhonationTime}. \tag{3.4}$$

There is another directly related feature, called **average syllable duration (ASD)**, which is the reciprocal of the articulation rate:

$$ASD = \frac{PhonationTime}{|SyllableNuclei|}. \tag{3.5}$$

3.2.3 Visual Features

The Linux command pdftotext is used to extract **text layout information**. It extracts the position of each text element and size of the slide. There is a hierarchical relationship between text elements: the biggest text element is text flow, which contains multiple text lines; each text line consists of multiple words. This information is converted as a XHTML file.

Similarily, the pdftohtml command is used to extract the image position and size of the slide as a XML file.

The generated XHTML and XML files are then parsed to JSON files, which are convenient for data handling.

For each page, given the layout information of texts, images, and the slide area, the area where texts or images overlap with slide is calculated. The **text ratio** is calculated as the ratio of overlapped text area to the whole slide area:

$$TextRatio = \frac{OverlappedTextArea}{SlideArea}. \tag{3.6}$$

Similarity, the **image ratio** is calculated as:

$$ImageRatio = \frac{OverlappedImageArea}{SlideArea}. \tag{3.7}$$

For the whole PDF file, the mean and sample variance of the text ratio and image ratio values are calculated and stored in JSON files.

3.3 Cross-modal Features

By far, the introduced features can represent a video only in a single modality. In order to analyse a video comprehensively, features which are generated from multiple modalities at the same time should be identified. These features are called cross-modal features. In this section, three cross-modal features are introduced: highlight of important statements, level of detailing, and coverage of slide contents.

During feature extraction, these natural language processing functions are used: noun selection and counting, lemmatisation, and finding synonyms. The functions are adapted from Bird et al. [3].

3.3.1 Highlight of Important Statements

Highlight of important statements can indicate how often the important statements in lecture are emphasised.

The feature requires two inputs: emphasised statements from speech and important statements from slides. The following two parts are about searching emphasised statements based on audio and transcription. The

important statements are extracted from text layout information, which is explained in the third part. Finally, the average highlight time for important statements is calculated (see formula 3.8).

Search Local Maximum Timestamp

If a speaker emphasises a statement, it is supposed that there is a corresponding local maximum in audio signal. In order to extract it, the speech is divided into multiple blocks sequentially in unit of ten seconds. Ten seconds is the same interval as the analysis for PVQ average (cf. description of PVQ average in subsection 3.2.1). And the local maximum block is extracted, if the signal value of that block is local maximum. The value for a block is calculated as an average value of existing extracted values for every one second.

The local maximum block is at first searched separately for these three signals (aspects): F_0, loudness, and energy. Then the local maximum blocks, which appear in results of three aspects at the same time, are selected. The timings of the extracted blocks are returned by program for locating the most possible transcription.

It is necessary to explain the calculation for aspect loudness. For loudness, the average value of loudness and modulated loudness is used to search local maxima. These two measurements are used in different situations, so the average of them is used to reduce noise in the signal (cf. description of loudness in subsection 3.2.1).

Locating Transcription

The local maximum timestamp represents an emphasis of the corresponding time block. In order to know what the speaker has said in that emphasis, it is needed to locate the transcription.

Transcriptions in my dataset are presented as SRT files. A SRT file contains a collection of subtitle units (SRT collection) which are numbered sequentially. Every subtitle unit (SRT unit) consists of a text with begin and end timestamp.

The SRT unit, the middle timestamp of which is nearest to local maximum timestamp, is selected as the most possible SRT unit (localMaxSrt in algorithm 2).

If there are multiple SRT units which has an overlap with the block, it is also possible that one of them represents the local maximum. Number of nouns (function NN(su) for SRT unit su) from transcriptions, is counted for overlapping SRT units and localMaxSrt. If there are two more nouns from a SRT unit than localMaxSrt, the nearest one is chosen to replace localMaxSrt. (see algorithm 2)

Algorithm 2: Locate Transcription

input : a local maximum block `localMaxBlock` with middle timestamp `t`, SRT collection `sc`
output: most possible corresponding SRT unit

```
1  surroundingSrtList ← empty list;
2  foreach srt unit su in sc do
       // phrases of relation conditions
       // (contains, is in, and has an overlap with)
       // means relations regarding timing
3      if su contains localMaxBlock then
4          localMaxSrt ← su;
5          break;
6      else
7          if su is in localMaxBlock then
8              localMaxSrt ← su;
9          else
10             if su has an overlap with localMaxBlock then
11                 append su to surroundingSrtList;
12             end
13         end
14     end
15 end
16 candidateSrtList ← empty list;
17 foreach SRT unit su in surroundingSrtList do
       // NN(su) returns the number of nouns
       //          in transcription from srt unit su
18     if NN(su) − NN(localMaxSrt) >= 2 then
19         append su to candidateSrtList;
20     end
21 end
22 if candidateSrtList is not an empty list then
23     find the su whose middle timestamp is nearest to t;
24     localMaxSrt ← su;
25 end
26 return localMaxSrt;
```

The transcriptions of all local maximum SRT units represent the emphasised statements from a speaker.

Important Statements Extraction

For each video, there is a corresponding PDF file available in my dataset. The important statements are extracted from it. At the same time, synonyms are also searched, which belong to the set of important statements as well.

The text layout information is already extracted and stored in a JSON file. Based on this information, the text lines are at first sorted by the area. Texts from first two biggest ones, which are supposed to contain important statements, are selected.

Words of nouns are searched from texts of selected lines. Synonyms are then searched for these nouns. All nouns and their synonyms are lemmatised as well. The lemmatisation can help to count the highlight more conveniently. The important statements are the lemmas of the nouns and their synonyms.

Calculation

As mentioned, important statements are represented as a list of lemmas (impStatList). The emphasised statements are lemmatised as well to a list (emphasisedContent). Lemmatisation function LEM(content) lemmatise content and returns a list with no duplicate lemmas. For each lemma from emphasisedContent, it is then checked whether it is in impStatList. If yes, count the highlight time. (see algorithm 3)

Highlight of Important Statements is calculated as the average highlight time for all lemmas from important statements:

$$Highlight = \frac{TotalHighLightTime}{|LEM(ImportantStatements)|}. \tag{3.8}$$

3.3.2 Level of Detailing

The level of detailing indicates how detailed a speaker has explained for each slide. The measurement is straightforward: the ratio of number of said words to number of words on the slide (see formula 3.10).

The said words are documented in transcriptions. For searching corresponding transcriptions for a slide, the video should be segmented according to slide content. So in each segmentation, the slide stays still.

Algorithm 3: Calculation of Average Highlight Time

input : list of emphasised SRT units `empSRTs`, list of lemmas from important statements `impStatList`
output: average highlight time

```
// LEM(empSRTs): lemmalise transcriptions from empSRTs,
//               return a list of lemmas, each lemma is unique in the list
emphasisedContents ← LEM(empSRTs);
time ← 0;
foreach lemma in empSRTs do
    if lemma is found in impStatList then
        time ← time + 1;
    end
end
len ← length of impStatList;
return time/ len;
```

Video Segmentation and Slide Matching

The content-based scene detection method is used for video segmentation. The program is adapted from PySceneDetect [5]. Fast changes or cuts in video content are detected. After that, the corresponding slide page number need to be mapped to each segmentation.

There are two possibilities for mapping: text matching and image descriptors matching. The principle is quite simple: find the page which has the most text or most image descriptors in common with the frame in the middle of a scene.

Among these two methods, the text matching is preferred. The image descriptors matching is then used under the following conditions: (a) no text is recognised from the frame; (b) the text from frame is not found in slides.

The text is extracted using Tesseract OCR [25]. Under the right conditions, the Scale Invariant Feature Transform (SIFT) descriptors, which are proposed from Lowe [21], are chosen as image descriptors for matching. The reason is that the SIFT descriptors are "invariant to image scale and rotation" (cf. Lowe [21]). It is suitable for matching from frame to slide. (see algorithm 4)

Figure 3.2 is a frame from a video in my dataset. A slightly skewed slide is appeared in the up-left side of the figure. The orginal slide is in rectangle. In this situation, most of the SIFT features from original slide remain in the video frame, so that a matching based on these features is possible. Therefore, the image descriptors matching can be used.

The whole process is described in algorithm 4.

Figure 3.2: A Frame Sample from Video 4_2a

Calculation

From a slide, the words are at first extracted. The corresponding video clips are found based on segmentation information.

For these clips, the number of said word is counted. Overlapped SRT units are to be found. If whole SRT unit is overlapped with a video clip, all words of that SRT unit will be counted; if just a part of SRT unit is overlapped with a video clip, the number of words is calculated regarding to the overlapping ratio to the whole SRT unit duration:

$$|SaidWords| = |WordsInSRT| \times \frac{OverlapDuration}{SRTUnitDuration}. \tag{3.9}$$

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

$$LevelOfDetailing = \frac{|SaidWords|}{|WordsOnSlide|}. \tag{3.10}$$

So the level of detailing values for each page are calculated. The mean and sample variance of these values are then calculated for later usage.

3.3.3 Coverage of Slide Contents

The coverage of slide contents indicates how much the slide content is covered. It is also a measurement per slide. For a slide, the set of words on it ($WordsOnSlide$) are extracted.

The corresponding transcriptions (SRT units) are found according to segmentation information, which is the same as in subsection 3.3.2. Given a

Algorithm 4: Video Segmentation and Slide Matching

Data: MP4 video file, PDF slide file
Result: Mappings of video clip to slide page number

```
use content-based scene detection for video;
foreach scene do
    frameImg ← frame at the middle timestamp of the scene;
    text ← OCR(frameImg);
    siftDescriptors ← SIFT(frameImg);
    if text is not empty then
        match text with the page which has the most common texts;
        if text is not found in all pages then
            match siftDescriptors with the page which has the
             most common sift descriptors;
        end
    else
        match siftDescriptors with the page which has the most
         common sift descriptors;
    end
end
```

slide, the corresponding video clips are found according to the slide number. Based on these video clips, the said words can be extracted from overlapping SRT units.

For a SRT unit which has a small overlap ratio, selection of said words in the transcript is very difficult. For simplicity, a SRT unit with the overlapping ratio which is bigger than 0.8, is selected. Then the set of said words ($WordsInSrt$) is generated from the complete transcription of selected SRT units.

The words in $WordsOnSlide$ and $WordsInSrt$ are all lemmatised. The **coverage of slide content** is calculated as the ratio of the number of common words in $WordsInSrt$ and $WordsOnSlide$ to number of words in $WordsOnSlide$:

$$Coverage = \frac{|WordsInSrt \cap WordsOnSlide|}{|WordsOnSlide|}. \tag{3.11}$$

Similarly, the mean and sample variance of the values of all pages are then calculated for later usage.

3.4 Summary of All Features

Table 3.1 summaries all features which are proposed in this chapter.

Table 3.1: List of All Features

Modality	**Features**
audio	F_0
	PVQ Average
	Loudness
	Modulated Loudness
	RMS Energy
	Absolute Period to Period Jitter
	δ Jitter
	Absolute Period to Period Shimmer
	Harmonicity (Spectral)
	Log. HNR
linguistic	Speech Rate
	Articulation Rate
	ASD
	Subtitles and Their Timeframes
visual	Text Layout Information
	Text Ratio*
	Image Ratio*
cross	Highlight of Important Statements
	Level of Detailing*
	Coverage of Slide Contents*

*: The feature is calculated per slide. For the whole PDF file, the mean and sample variance of the values are calculated.

Chapter 4

Evaluation

This chapter introduces the design of the user study.

One purpose of this user study is to find the correlation between features and quality of the video. This study therefore asks many questions about the quality in multiple aspects. The usefulness of extracted features can then be proved. Furthermore, the collected user opinion can serve as input for training a scoring model.

The content of this chapter is organised as follows: at first, the organisation of all videos and the individual video set for each subject are introduced; the experiment operation is then explained; specifically, the criteria in evaluation form are considered with regards to modalities; the calculation of the knowledge gain is explained in the last section.

4.1 Evaluation Organisation

For annotation, a set of twenty-three videos are organised. The subject of the videos is software engineering.

There is not more than one presenter for each video. Up to four of the videos have the same presenter. Each video is about five to ten minutes long. The complete video list is attached in appendix A.

The subjects are thirteen adults with a computer science background. Each subject should see nine videos. The set is organised to contain as many presenters as possible. The distribution for the number of presenters which are seen by each subject is shown in figure 4.1. Each video is viewed minimally five times by different subjects.

4.2 Experiment

Before seeing each video, the subject is asked to answer five questions about the video. It is supposed that the subject can concentrate more on videos, after answering these questions.

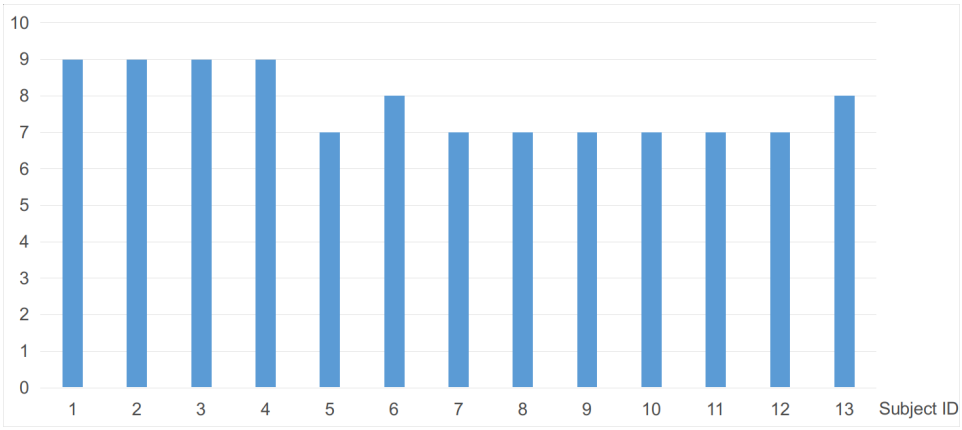

Figure 4.1: Distribution for Number of Presenters

With the goal of setting a difficult question list for each video, the following conditions are set: (a) In the question list for a video, only two to four questions are related to that video; (b) For each question there is at least one correct answer.

After filling out the first form, the subject should see the video till the end. During this time, the subject may not take notes or pause the video.

The subject should fill in the same question form again, after seeing the entire video. At last, the evaluation form will be filled in.

The evaluation form is designed to give a feedback on each seen video with regards to multiple aspects. The content of the form will be explained in the next section.

4.3 Criteria for Evaluation

All modalities, which are mentioned in chapter 3, are asked in evaluation form. The corresponding criteria in the form are shown in table 4.1.

Each criterion in the evaluation form is to be scored on a scale from one to five. The bigger the score, the better the performance for the criterion. The evaluation form is attached in appendix B.4.

What's more, the knowledge gain is also calculated, according to the given answers before and after seeing a video.

Let $ScoreBefore$ be score before seeing the video, $ScoreAfter$ be score after seeing the video, $FullScore$ be score for totally correct answer, knowledge gain is calculated as:

$$KnowledgeGain = \frac{ScoreAfter - ScoreBefore}{FullScore}$$

The calculation of score for each question is explained as follows. Firstly, the score for an empty answer will be calculated as zero. The reason for it

Table 4.1: Modalities and Their Corresponding Items

Modality	**Corresponding Items in Evaluation Form**
audio	clear language vocal diversity
linguistic	filler words speed of presentation
visual	text/image/formula/table design structuring of the presentation
cross	coverage of the slide content appropriate level of detail highlight of important content summary overall rating

is that subjects are told to leave the answer row free if they do not know the answer (see appendix B.3.1). If the answer row is not empty, the score will then be calculated. Each selection of a question is rated. It should be checked or unchecked in the correct answer. For that selection, if a subject has given the correct answer, the score will increase by one; if not, decrease by one. The initial score for a question is zero.

Say for a question there are four selections: A, B, C, and D. The correct answer is: A checked, B checked, C checked, D unchecked. So $FullScore$ is four, the number of selections. If a subject has selected only A and B, then the answer is: A checked (+1), B checked (+1), C unchecked (−1), D unchecked (+1). So the score for the given answer is two.

Chapter 5

Discussion

This section analyses the correlation between generated features and evaluation results.

At first, a general overview of the correlation selection and data preprocessing for evaluation results is introduced. As correlations of cross-modal features are weak for automatic segmentation, the videos are manually segmented to find if correlations change. The correlation difference has shown that manual segmentation delivers stronger results.

The next sections discuss positive and negative correlations. The possible reasons are given as well. For cross-modal features, the correlations are more deeply investigated.

At last, the correlations with knowledge gain indicate encouraging and promising results.

5.1 Correlation Analysis

It is important to look at the data types in measurement scales at first for a selection of correlation. After the calculation, the correlation results need to be interpreted according to the correlation method and the size of samples.

5.1.1 Data Measurement Scales

All generated features are interval data because the order and exact differences can be inferred from the values. The data[1] which are collected from evaluation form are ordinal data, which are scaled from one to five. Knowledge gain, which is calculated from control questions, is interval data.

For interval data, the Pearson correlation is used. For ordinal data, the Spearman correlation is used.

For correlations between features and knowledge gain, the Pearson correlation has to be applied. For each video, the average knowledge gain is

[1] except the item "entry level", which is used for the project SALIENT [14]

calculated for all evaluators.

The Spearman correlation is used to analyse the correlations between features and evaluation results except knowledge gain. As the values from features are interval data, they are converted to a distinct rank as ordinal data. For each video, there are at least five evaluators who have given different scores. So the median of the scores, which is the point on the scale that divides the distribution into halves, is extracted for each evaluation question. There are twenty-two videos for correlation analysis, which means there are also twenty-two medians for each evaluation question.

5.1.2 Correlation Result Explanation

The correlation result is represented as correlation coefficient r and directional alpha level α. r ranges from -1 to $+1$. -1 means strong negative correlation, 0 means no correlation, and $+1$ means strong positive correlation. α means the possiblity to reject null hypothesis, i.e., $r = 0$. For example, $\alpha = 0.025$ can be interpreted as: there is less than 2.5% chance that the correlation happened by chance. The smaller α is, the more possible it is that there exists a correlation.

Each r corresponds to an α, which is dependent on the type of analysis and the size of samples. For samples with size N, and the degrees of freedom $df = N-2$, the table of exact critical values for Spearman from Ramsey [23], and for Pearson from Weathington et al. [27] (p. 452) is referenced in my thesis. As mentioned, there are twenty-two videos for correlation analysis (one from the annotation video set is excluded due to lack of PDF file). So for all following analyses, $N = 22$, $df = 20$.

5.2 Manual vs. Automatic Segmentation

In this section, these two cross-modal features are discussed: level of detailing and coverage of slide contents. These features are calculated according to the video segmentation and slide matching (cf. subsection 3.3.2).

The correlations between these features and evaluation results regarding to detailing and coverage are very weak. The reason may lie in imprecise segmentation information.

Figure 5.1 has shown the results of a video clip in which three successive slides are shown. The correct segmentation is shown above the arrow in black text. For automatic segmentation (shown under the arrow as red text), there is only two segmentations. And the corresponding segmentation to the second slide is missing. Values of these features are zero for the second slide. This affects the correctness of these two cross-modal measurements.

Therefore, evaluation videos are manually segmented. After that, corresponding slide number is matched to each segmentation. The level

Figure 5.1: Manual and Automatic Segmentation

Table 5.1: Changes of Correlations

Features	**Evaluation Results**	$r_s^{'}$	r_s
Level of Detailing Mean	Clear Language	0.178	0.541
	Vocal Diversity	0.354	0.615
Level of Detailing Var.	Clear Language	0.219	0.448
	Appropriate Level of Detail	0.162	0.378
Coverage of Slide Contents Mean	Filler Words	0.391	-0.304
Coverage of Slide Contents Var.	Appropriate Level of Detail	-0.230	-0.524

$r_s^{'}$: Spearman Correlation Coefficient generated from automatic segmentation
r_s: Spearman Correlation Coefficient generated from manual segmentation

of detailing and the coverage of slide contents are calculated again based on manually generated segmentation information.

Table 5.1 has shown a significant change in correlation value between these features and all listed evaluation results (the absolute difference value is greater than 0.2).

Focusing on the cross-modal feature from the evaluation results, i.e., appropriate level of detail (hereinafter referred to as detail), it is expected that correlations exist. But after automatic segmentation, the correlation coefficient $r_s^{'}$ between *level of detailing* and *detail* is small (0.162), while the correlation coefficient r_s is greater (0.378) after manual segmentation.

The results from automatic and manual segmentation have shown that a correct segmentation can help find correlations for features level of detailing and coverage of slide contents. Hence, the following correlation analyses are based on the manually generated segmentation information.

5.3 Positive Spearman Correlations

Table 5.2 has selected positive Spearman correlation with α which is less than 0.05. The following subsections are going to explain correlations in different modalities.

Table 5.2: Positive Spearman Correlations

Features	**Evaluation Results**	r_s	$\alpha <$
Loudness	Clear Language	0.430	0.025
RMS Energy	Clear Language	0.368	0.05
F_0	Clear Language	0.608	0.0025
	Vocal Diversity	0.429	0.025
	Summary	0.420	0.05
Absolute Period to Period Jitter	Speed of Presentation	0.423	0.05
Delta Jitter	Speed of Presentation	0.408	0.05
Harmonicity (spectral)	Clear Language	0.435	0.025
Log. HNR	Vocal Diversity	0.389	0.05
Speech Rate	Vocal Diversity	0.525	0.01
Articulation Rate	Filler Words	0.459	0.025
ASD	Image Design	0.371	0.05
Image Ratio Mean	Filler Words	0.394	0.05
Highlight of Important Statements	Coverage of the slide content	0.471	0.025
Level of Detailing Mean	Clear Language	0.541	0.01
	Vocal Diversity	0.615	0.0025
	Filler Words	0.601	0.0025
Level of Detailing Var.	Clear Language	0.448	0.025
	Vocal Diversity	0.463	0.025
	Filler Words	0.493	0.025
	Appropriate Level of Detail	0.378	0.05

r_s: Spearman Correlation Coefficient
α: Directional Alpha Level

5.3.1 Audio Features

These features have shown a strong correlation ($\alpha < 0.025$) with *clear language*: *loudness*, *RMS energy*, and *harmonicity (spectral)*. According to subsection 3.2.1, loudness and RMS energy represent the power or strength in the voice, harmonicity (spectral) indicates the signal quality: all of them have a relationship with the clarity of the language.

On the other hand, another measurement of quality *log. HNR* has a relationship with *vocal diversity*. It is encouraging to compare this correlation to the result of the study from Yumoto et al [30], which indicates a negative correlation between HNR and hoarseness.

F_0 has a strong correlation with two audio modal aspects from evaluation: *clear language* ($\alpha < 0.0025$) and *vocal diversity* ($\alpha < 0.025$). What's more, F_0 has a moderate correlation ($\alpha < 0.05$) with cross-modal aspect *summary*.

Above findings about audio features broadly support that the mentioned features can have an effect on the aural perception.

Surprisingly, jitter-related features were found to have a correlation with the *speed of presentation*.

5.3.2 Linguistic and Visual Features

For the *articulation rate*, there is a strong correlation ($\alpha < 0.025$) to *filler words*. The observed correlation indicates that there is a strong possibility that these situations are existing together: (a) lots of filler words are said; and (b) many syllables are said in unit speaking time.

The *speech rate* has a correlation to linguistic aspect *vocal diversity*.

There is a correlation between *ASD* and *image design*. It is therefore likely that these three events are happening at the same time: (a) the user feels there are too many images on the slide; (b) the text on the slide is very short; and (c) the speaker has spoken for a short time.

One unanticipated finding was that there is a correlation between *image ratio mean* and *filler words*.

5.3.3 Cross-modal Features

The correlations to cross-modal aspects in evaluation results are at first discussed. It is observed that *highlight of important statements* correlates with *coverage of the slide content*. *Level of detailing variance* is correlated with *appropriate level of detail*. These findings have shown that the proposed cross-modal features can represent the cross-modal perception.

The mean and sample variance of *level of detailing* have a correlation with *filler words*. This relationship may partly be explained by the nervousness of the speaker. If the speaker is nervous, it would be likely that the content for each page is explained either too much or not enough, with lots of filler words.

For two *level of detailing* related features, there exist correlations to *clear language* and *vocal diversity*. Level of detailing is from linguistic and visual modality, while clear language and vocal diversity come from audio modality. This finding was unexpected and suggests that the perceptions from all modalities influence each other.

5.4 Negative Spearman Correlations

Table 5.3 has selected negative Spearman correlation with α which is less than 0.05. The following subsections are going to explain correlations in different modalities.

Table 5.3: Negative Spearman Correlations

Features	**Evaluation Results**	r_s	$\alpha <$
Absolute Period to Period Shimmer	Clear Language	-0.623	0.0025
	Vocal Diversity	-0.437	0.025
PVQ Average	Clear Language	-0.416	0.05
Speech Rate	Overall Rating	-0.455	0.025
Articulation Rate	Image Design	-0.368	0.05
ASD	Filler Words	-0.459	0.025
Text Ratio Mean	Vocal Diversity	-0.396	0.05
Text Ratio Var.	Speed of Presentation	-0.454	0.025
Image Ratio Mean	Text Design	-0.466	0.025
Image Ratio Var.	Speed of Presentation	-0.561	0.005
Coverage of Slide Contents Var.	Clear Language	-0.454	0.025
	Appropriate Level of Detail	-0.524	0.01
	Summary	-0.434	0.025

r_s: Spearman Correlation Coefficient
α: Directional Alpha Level

5.4.1 Audio Features

The *absolute period to period shimmer* (shimmer) is at first discussed. There is a strong negative correlation ($\alpha < 0.0025$) with *clear language*. It correlates with *vocal diversity* ($\alpha < 0.025$) as well. This finding confirms that shimmer is an indicator about sickness in the voice (cf. subsection 3.2.1), which affects the performance in language.

For *PVQ average*, it has a negative correlation with *clear language*. This correlation has been unable to demonstrate that higher PVQ mirrors more liveliness according to Hincks [18]. This inconsistency may be due to the data smoothing for F_0 values. The smoothed values are no more accurate and the error values could still exist after smoothing.

5.4.2 Linguistic Features

Speech rate has a negative correlation with *overall rating*. This observation may support the hypothesis that fast speaker has bad notes. But a correlation between *speech rate* and *speed of presentation* is not found. This

missing correlation may be due to subjective evaluation. It is important to bear in mind the possible bias in answers regarding speed.

ASD has a negative correlation to *filler words*, which corresponds to the positive correlation between *articulation rate* and *filler words*, since ASD is the reciprocal of the articulation rate. The same principle applies to the negative correlation between *articulation rate* and *image design*.

5.4.3 Visual Features

The observed strong negative correlation ($\alpha < 0.025$) between *image ratio mean* and *text design* might be explained in this way: a slide is full of images and there is little text on it.

The other correlations to evaluation results of audio and linguistic modality are also listed, which are not intuitive. And the correlations from *text ratio* to *text design* and from *image ratio* to *image design* are very weak. This result may be explained by the fact that most slides from my dataset are full of images and with little text. There are not enough data with a great variety. Another possible explanation for this is that ratio information could not reflect the design.

5.4.4 Cross-modal Features

Coverage of slide contents variance has a strong negative correlation ($\alpha < 0.01$) with *appropriate level of detail*. According to the result, it could be suggested that the lecture content is not properly explained in detail, because there is a significant difference of the content coverage across all slides.

Coverage of slide contents variance has also a negative correlation with *summary*. This relationship may partly be explained as: the lecture content for each page is either fully summarised or not at all summarised.

The reason for the correlation with *clear language* is not coherent but it may exist a correlation between linguistic and audio modality, because the coverage of slide contents is generated from linguistic (and visual) modality and clear language comes from audio modality.

5.5 Correlations of Cross-modal Features

Specifically, correlations between cross-modal features and evaluation results are discussed.

It is expected that there is a correlation between *highlight of important statements* from feature set and *highlight of important statement* from evaluation results. But it is observed that *highlight of important statements* has a relationship with *coverage of the slide content*. A possible reason lies

in extraction of important statements. Important statements are a complete sentence or phrase. In subsection 3.3.1, they are lemmatised for simplicity.

The correlation for coverage is also very weak from extracted features to evaluation results. Instead, *coverage of slide content variance* from feature set negatively correlates with *appropriate level of detail* from evaluation results. This inconsistency may due to a subjective evaluation. It is very difficult for a user to give an answer about "appropriate level".

5.6 Correlations with Knowledge Gain

Table 5.4 has shown correlations between features and knowledge gain. The correlations, whose $\alpha < 0.1$, are selected.

Table 5.4: Correlations to Knowledge Gain

Features	r_p	$\alpha <$
Modulated Loudness	-0.483	0.025
Highlight of Important Statements	0.398	0.05
Harmonicity (spectral)	0.323	0.1
RMS Energy	0.299	0.1

r_p: Pearson Correlation Coefficient
α: Directional Alpha Level

The most obvious finding to emerge from the analysis is that there is a strong negative correlation ($r_p = -0.483, \alpha < 0.025$) from *modulated loudness*. This result may be explained by the fact that loud noise disturbs the learning process.

Contrary to that, *harmonicity (spectral)*, and *RMS Energy* positively correlate with *knowledge gain*. It may be that these participants have felt comfortable hearing a pleasantly sonorous sound.

It is interesting to note that *highlight of important statements* has a positive correlation with knowledge gain. A possible explanation for this might be that the highlighting can help to memorise an important statement.

Chapter 6

Conclusion and Outlook

This thesis has explored the way to assess video quality automatically. Lecture videos are chosen for feature extraction. Multimodal features are generated not only from single modality, but also from multiple ones. Three cross-modal features are first proposed as measurements for teaching performance of speaker: (a) the highlight of important statements (audio, linguistic, and visual modality), (b) the level of detailing (linguistic and visual modality), and (c) the coverage of slide contents (linguistic and visual modality).

A user study has been conducted to investigate if the proposed features can represent a video in multiple aspects. According to the results of the user study, most of the features fit this requirement. Features except text ratio mean and sample variance, image ratio sample variance, and the coverage of slide contents sample variance, have shown correlations to users' opinions for a video.

Furthermore, the correlation between features to learning effect is studied as well. There are not only positive but also negative correlations. For cross-modal feature highlight of important statements, there is a strong correlation to knowledge gain, which indicates this feature may enhance learning.

All these findings contribute in several ways to our understanding of learning on the web and further research on the improvement of it.

There is still a lot of work to do. Firstly, the negative correlation of *PVQ average* with *clear language* should be further investigated. Besides, features about the design of text and image need to be improved.

Secondly, chapter 5 has shown automatic segmentation leads to inaccuracy in calculation for the level of detailing and the coverage of slide contents. The lecture video should be automatically segmented more precisely, in order to deliver more correct mapping information of video clip to slide number.

Thirdly, *highlight of important statements* from feature set do not have a strong correlation with *highlight of important content* from evaluation results. A possible reason about the extraction of important statements

is suggested. A more robust and accurate calculation should be researched to describe highlight better.

With all these steps completed, the quality of the lecture video can be assessed with all proposed features. This can be realised through machine learning methods. A video classifier can be trained using features and user evaluations. A video can then be classified to a quality category, which serves as a measurement for relevance in modern search engines.

Acknowledgements

I would like to express my gratitude to Professor Ralph Ewerth, my research supervisors Mr. Otto Christian and Dr. Anett Hoppe, for their patient guidance, enthusiastic encouragement, quick feedback, and useful critiques of this thesis. I would also like to thank Ms. Jasmin Farsi, for her valuable advice and professional assistance in writing my thesis. My grateful thanks are also extended to all participants for their help in the user study and Mr. Eric Müller-Budack who has helped me with the organisation of it.

I would also like to extend my thanks to Mr. Matthias Springstein for his assistance in helping me setting up my workstation in the laboratory of the visual analytics research group.

Finally, I wish to thank my parents and my girlfriend for their support and encouragement throughout my master thesis.

Appendix A

Video List

The video list for feature extraction and evaluation is attached in this appendix.

Videos for Evaluation

The videos are from course Globally Distributed Software Engineering which is hosted on edX [1].

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 [7]. A copy of the license is included in appendix C.

The dataset for this thesis was created by and adapted from material posted on the Delftx website, delftx.tudelft.nl, which was created by TU Delft faculty member Prof. Rini van Solingen, 2018. DelftX is not responsible for any changes made to the original materials posted on its website and any such changes are the sole responsibility of Jianwei Shi.

The first number of the ID means the week number, e.g. 1_2a means the first week.

[1] `https://courses.edx.org/courses/course-v1:DelftX+GSE101x+1T2018/course/`

Table A.1: Videos and Their Corresponding Items

Video ID	Course Title	Presenter
1_2a	What is GDSE?	Rini van Solingen
1_2b	Why do GDSE?	
1_2c	Cultural Differences	
1_2d	GDSE Research	
1_3a	GDSE at Exact Online - Process and Tools	Emiel Romein
1_3b	GDSE at Exact Online - Product	
1_3c	GDSE at Exact Online - People	
2_2a	Lean	Eelco Rustenburg
2_2b	Scrum	
2_2c	The Agile Manifesto	
2_2d	Large Scale Agile	
2_3a	Examples of Distributed Scrum	Jeff Sutherland
3_2b	Some Do's and Don'ts while Doing Automation	Prajeesh Pratap
3_3a	Automation of a CD pipeline	Erik Ammerlaan
3_3b	Continuous Delivery at Exact	
4_2a	Outsourcing from a Decision-maker Perspective	Suzanne Kelder
4_3a	Top 5 Lessons Learned for Selecting a Near- or Offshore Vendor	Svenja de Vos
5_1a	Offshoring and Cost Savings	Darja Šmite
5_1b	Bottom-line cost of offshoring in "SwedCo"	
6_2a	Culture in Global Software Engineering	Dianne Elsinga
7_2b	Tools for a Distributed Software Engineer	Marudhamaran Gunasekaran
7_3a	Developments in Country Selection	Paul Tjia
7_3c	Palestine and North Korea	

Appendix B

Evaluation Organisation

This appendix gives detailed information about the design and organisation for the evaluation. Forms for answering questions and evaluation are attached at the end.

B.1 Subjects Information

This section introduces the information about subjects. In total, thirteen students have taken part in the evaluation. All of them have a computer science background. They are between the ages of twenty-two and thirty, the distribution of which is shown in figure B.1. Eleven subjects have a Bachelor's Degree, while one has a Master's Degree and another one is now studying Bachelor (see figure B.2 left side). There are three female students in the whole group (see figure B.2 right side).

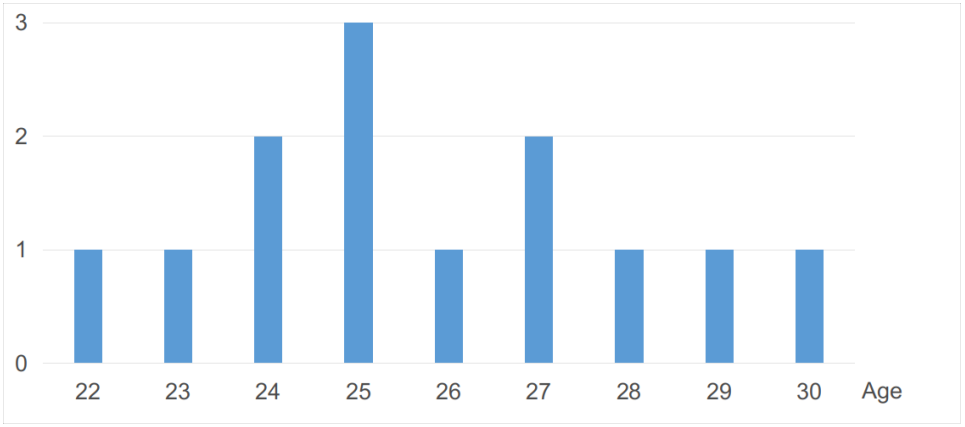

Figure B.1: Distribution for Age of Subjects

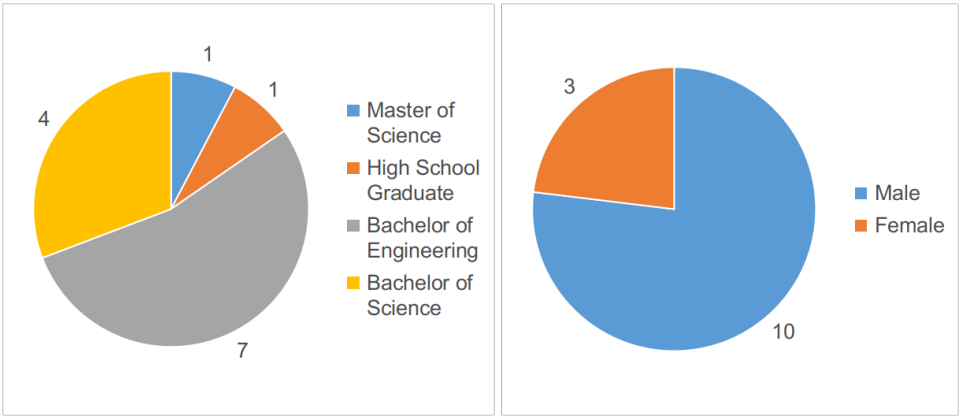

Figure B.2: Degree and Gender Information of Subjects

B.2 Design

The evaluation is conducted in a software. Different stimuli, such as text, video, and so on, can be designed and ordered.

In my user study, a design for a single video is shown in figure B.3.

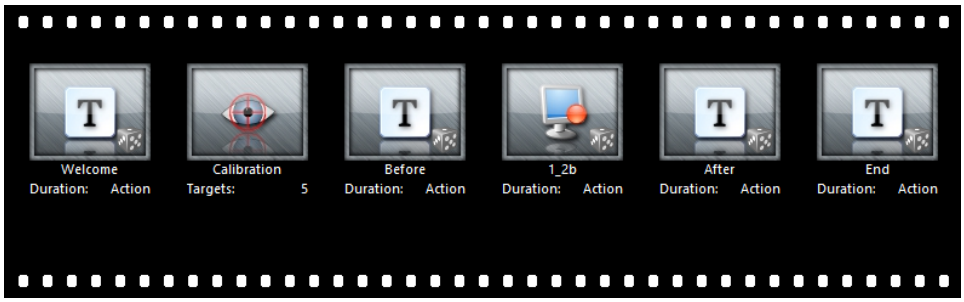

Figure B.3: Design for a single video

The procedure for a single video is explained by the following steps:

1. Welcome (introduction) text is shown.
2. Camera for eye tracking is calibrated.
3. Instruction text before seeing the video is shown, let a subject answer questions.
4. The subject clicks corresponding video and watches it. Eye tracking data is collected.
5. Instruction after seeing the video is shown, let the subject answer questions again and fill in evaluation form.

6. Software shows text indicating the study has come to an end.

For more videos, the steps from three to five are repeated with different video ID.

B.3 Instruction text

This section shows text which is used in the study. The following text is for the video with ID 1_2b.

B.3.1 Introduction

Welcome to the user study.
This introduction gives a brief outline of the upcoming procedure. Please read it carefully to ensure a successful user study.
Step 1. A short calibration process to adjust the eye-tracker to your height and distance from the PC. Follow the red dot with your eyes until the calibration is done.
Step 2. Fill out a short user data form asking for your age, level of education and age. This data will be anonymised.
Step 3. Afterwards the textual instructions for the actual experiment start.
General guideline for the experiment: Before watching a video, answer the provided questionnaire with the corresponding video ID.
After watching **the entire** video, answer the same questionnaire again and fill out the evaluation form.
Notice:
- It is not allowed to take notes during video playback.
- Only answers filled out with a pen will be counted.
- There is at least one correct answer for each question.
- Please leave the answer row free if you do not know the answer (instead of guessing).
- There will be nine videos to annotate. Take a break in-between videos if you want to.
Have fun ;-)

B.3.2 Before Playing Video

Fill in the control question form before seeing the video **1_2b**.
After clicking OK, open it from desktop.
After watching the video, click OK on the right down corner.

B.3.3 After Playing Video

Fill in the
1. control question form

2. evaluation form
after seeing the video **1__2b**.
You may take a break before seeing next video.
When you are ready, click OK.

B.4 Forms

Control question form (shown as Control Questions - Answering Card) and evaluation form are attached.

Control Questions - Answering Card

Video ID:		For Conductors only		
		Points for form A	Points for form B	Difference (B - A)
Person ID:				

Introduction

This study is part of the master thesis *Automatic Quality Assessment of Lecture Videos Using Multimodal Features*. For the evaluation of the thesis, a set of manually generated labels is necessary, which we want to generate in this study. Each participant will annotate nine videos and the study will take about two and a half hours.

Before watching the video

Form A

Questions	A	B	C	D	E	F	G	H	I
1	☐	☐	☐	☐	☐	☐	☐	☐	☐
2	☐	☐	☐	☐	☐	☐	☐	☐	☐
3	☐	☐	☐	☐	☐	☐	☐	☐	☐
4	☐	☐	☐	☐	☐	☐	☐	☐	☐
5	☐	☐	☐	☐	☐	☐	☐	☐	☐

After watching the video

Form B

Questions	A	B	C	D	E	F	G	H	I
1	☐	☐	☐	☐	☐	☐	☐	☐	☐
2	☐	☐	☐	☐	☐	☐	☐	☐	☐
3	☐	☐	☐	☐	☐	☐	☐	☐	☐
4	☐	☐	☐	☐	☐	☐	☐	☐	☐
5	☐	☐	☐	☐	☐	☐	☐	☐	☐

Please turn around the page and fill out the evaluation form.

Evaluation Form

Video ID:		
Person ID:		

1	2	3	4	5
is not true	is rarely true	is sometimes true	is mostly true	is absolutely true

Clear language: The spoken language is easy to understand.

☐ ☐ ☐ ☐ ☐

Vocal diversity: The use of variations in tone, tempo and volume is good.

☐ ☐ ☐ ☐ ☐

Filler words: The lecturer hardly used any filler words (ahh, ehh, and, ...).

☐ ☐ ☐ ☐ ☐

Speed of presentation: The speed of presentation is appropriate.

☐ ☐ ☐ ☐ ☐

Coverage of the slide content: The lecturer considers the entire content of the slide.

☐ ☐ ☐ ☐ ☐

Appropriate level of detail: The presenter explains the content in detail, if necessary.

☐ ☐ ☐ ☐ ☐

Highlight of important content: The presenter highlights the important content.

☐ ☐ ☐ ☐ ☐

Summary: The lecturer summarises the learning content frequently.

☐ ☐ ☐ ☐ ☐

Design of materials (presentation slides/whiteboard/flipchart):

Text: The amount of text per slide is appropriate. (Please leave blank if there is no text)

☐ ☐ ☐ ☐ ☐

Image: The amount of images per slide is appropriate. (Please leave blank if there are no images)

☐ ☐ ☐ ☐ ☐

Formula: The amount of formulas per slide is appropriate. (Please leave blank if there are no formulas)

☐ ☐ ☐ ☐ ☐

Table: The amount of tables per slide is appropriate. (Please leave blank if there are no tables)

☐ ☐ ☐ ☐ ☐

Structuring of the presentation: The presentation is well structured.

☐ ☐ ☐ ☐ ☐

Entry level: For which target group is the video suitable?

Beginners		Advanced Learners		Experts
☐		☐		☐

Overall rating: Overall, I rate the training video as

very bad	bad	average	good	very good
☐	☐	☐	☐	☐

Appendix C

License

Attribution-NonCommercial-ShareAlike 4.0 International

==========

Creative Commons Corporation ("Creative Commons") is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an "as-is" basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Using Creative Commons Public Licenses

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

```
Considerations for licensors: Our public licenses are
intended for use by those authorized to give the public
permission to use material in ways otherwise restricted by
copyright and certain other rights. Our licenses are
irrevocable. Licensors should read and understand the terms
and conditions of the license they choose before applying it.
Licensors should also secure all rights necessary before
applying our licenses so that the public can reuse the
material as expected. Licensors should clearly mark any
material not subject to the license. This includes other CC-
licensed material, or material used under an exception or
limitation to copyright. More considerations for licensors:
wiki.creativecommons.org/Considerations_for_licensors
```

Considerations for the public: By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor's permission is not necessary for any reason--for example, because of any applicable exception or limitation to copyright--then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable. More considerations for the public:
wiki.creativecommons.org/Considerations_for_licensees

==========

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

a. Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.

b. Adapter's License means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.

c. BY-NC-SA Compatible License means a license listed at creativecommons.org/compatiblelicenses, approved by Creative Commons as essentially the equivalent of this Public License.

d. Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.

e. Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.

f. Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.

g. License Elements means the license attributes listed in the name of a Creative Commons Public License. The License Elements of this Public License are Attribution, NonCommercial, and ShareAlike.

h. Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied this Public License.

i. Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.

j. Licensor means the individual(s) or entity(ies) granting rights under this Public License.

k. NonCommercial means not primarily intended for or directed towards commercial advantage or monetary compensation. For purposes of this Public License, the exchange of the Licensed Material for other material subject to Copyright and Similar Rights by digital file-sharing or similar means is NonCommercial provided there is no payment of monetary compensation in connection with the exchange.

l. Share means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material

available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.

m. Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.

n. You means the individual or entity exercising the Licensed Rights under this Public License. Your has a corresponding meaning.

Section 2 – Scope.

a. License grant.

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:

   `a. reproduce and Share the Licensed Material, in whole or in part, for NonCommercial purposes only; and`

   `b. produce, reproduce, and Share Adapted Material for NonCommercial purposes only.`

2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.
3. Term. The term of this Public License is specified in Section 6(a).
4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)

   (4) never produces Adapted Material.

5. Downstream recipients.

   a. Offer from the Licensor -- Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.

   b. Additional offer from the Licensor -- Adapted Material. Every recipient of Adapted Material from You automatically receives an offer from the Licensor to exercise the Licensed Rights in the Adapted Material under the conditions of the Adapter's License You apply.

   c. No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.

6. No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i).

b. Other rights.

1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.
2. Patent and trademark rights are not licensed under this Public License.
3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties, including when the Licensed Material is used other than for NonCommercial purposes.

Section 3 – License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. Attribution.

1. If You Share the Licensed Material (including in modified form), You must:

```
a. retain the following if it is supplied by the Licensor
with the Licensed Material:

i. identification of the creator(s) of the Licensed
Material and any others designated to receive
attribution, in any reasonable manner requested by
the Licensor (including by pseudonym if
designated);

ii. a copyright notice;

iii. a notice that refers to this Public License;

iv. a notice that refers to the disclaimer of
warranties;

v. a URI or hyperlink to the Licensed Material to the
extent reasonably practicable;

b. indicate if You modified the Licensed Material and
retain an indication of any previous modifications; and

c. indicate the Licensed Material is licensed under this
Public License, and include the text of, or the URI or
hyperlink to, this Public License.
```

2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.
3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.

b. ShareAlike.

In addition to the conditions in Section 3(a), if You Share Adapted Material You produce, the following conditions also apply.

1. The Adapter's License You apply must be a Creative Commons license with the same License Elements, this version or later, or a BY-NC-SA Compatible License.
2. You must include the text of, or the URI or hyperlink to, the Adapter's License You apply. You may satisfy this condition in any reasonable manner based on the medium, means, and context in which You Share Adapted Material.
3. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, Adapted Material that restrict exercise of the rights granted under the Adapter's License You apply.

Section 4 – Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database for NonCommercial purposes only;

b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material, including for purposes of Section 3(b); and

c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 – Disclaimer of Warranties and Limitation of Liability.

a. UNLESS OTHERWISE SEPARATELY UNDERTAKEN BY THE LICENSOR, TO THE EXTENT POSSIBLE, THE LICENSOR OFFERS THE LICENSED MATERIAL AS-IS AND AS-AVAILABLE, AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE LICENSED MATERIAL, WHETHER EXPRESS, IMPLIED, STATUTORY, OR OTHER. THIS INCLUDES, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR

A PARTICULAR PURPOSE, NON-INFRINGEMENT, ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OR ABSENCE OF ERRORS, WHETHER OR NOT KNOWN OR DISCOVERABLE. WHERE DISCLAIMERS OF WARRANTIES ARE NOT ALLOWED IN FULL OR IN PART, THIS DISCLAIMER MAY NOT APPLY TO YOU.

b. TO THE EXTENT POSSIBLE, IN NO EVENT WILL THE LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY (INCLUDING, WITHOUT LIMITATION, NEGLIGENCE) OR OTHERWISE FOR ANY DIRECT, SPECIAL, INDIRECT, INCIDENTAL, CONSEQUENTIAL, PUNITIVE, EXEMPLARY, OR OTHER LOSSES, COSTS, EXPENSES, OR DAMAGES ARISING OUT OF THIS PUBLIC LICENSE OR USE OF THE LICENSED MATERIAL, EVEN IF THE LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH LOSSES, COSTS, EXPENSES, OR DAMAGES. WHERE A LIMITATION OF LIABILITY IS NOT ALLOWED IN FULL OR IN PART, THIS LIMITATION MAY NOT APPLY TO YOU.

c. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 – Term and Termination.

a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.

b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

   1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or
   2. upon express reinstatement by the Licensor.

   For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.

d. Sections 1, 5, 6, 7, and 8 survive termination of this Public License.

Section 7 – Other Terms and Conditions.

a. The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.

b. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.

Section 8 – Interpretation.

a. For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.

b. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.

c. No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor.

d. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

==========

Creative Commons is not a party to its public licenses. Notwithstanding, Creative Commons may elect to apply one of its public licenses to material it publishes and in those instances will be considered the "Licensor." The text of the Creative Commons public licenses is dedicated to the public domain under the CC0 Public Domain Dedication. Except for the limited purpose of indicating that material is shared under a Creative Commons public license or as otherwise permitted by the Creative Commons policies published at creativecommons.org/policies, Creative Commons does not authorize the use of the trademark "Creative Commons" or any other trademark or logo of Creative Commons without its prior written consent including, without limitation, in connection with any unauthorized modifications to any of its public licenses or any other arrangements, understandings, or agreements

concerning use of licensed material. For the avoidance of doubt, this paragraph does not form part of the public licenses.

Creative Commons may be contacted at creativecommons.org.

Bibliography

[1] V. Balasubramanian, S. G. Doraisamy, and N. K. Kanakarajan. A multimodal approach for extracting content descriptive metadata from lecture videos. *Journal of Intelligent Information Systems*, 46(1):121–145, 2016.

[2] I. Bauer. *Moderne Ranking-Verfahren im WWW*. VDM Verlag Dr. Müller, 2007.

[3] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer [Computer program]. `http://www.fon.hum.uva.nl/praat`, 2018. [Version 6.0.37; retrieved 15-January-2019].

[5] B. Castellano. Video Scene Cut Detection and Analysis Tool [Computer program]. `https://github.com/Breakthrough/PySceneDetect#video-scene-cut-detection-and-analysis-tool`, 2018. [Version 0.5; retrieved 18-May-2019].

[6] L. Chen, C. W. Leong, G. Feng, and C. M. Lee. Using Multimodal Cues to Analyze MLA'14 Oral Presentation Quality Corpus: Presentation Delivery and Slides Quality. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, MLA, pages 45–52, New York, NY, USA, 2014. ACM.

[7] Creative Commons. Attribution-NonCommercial-ShareAlike 4.0 International. `https://creativecommons.org/licenses/by-nc-sa/4.0`, 2019. [Online; accessed 18-April-2019].

[8] N. H. de Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, May 2009.

[9] N. H. de Jong and T. Wempe. Praat Script Syllable Nuclei v2. `https://sites.google.com/site/speechrate/Home/`

praat-script-syllable-nuclei-v2, 2010. [Online; accessed 31-March-2019].

[10] F. Eyben. *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer, 2016. Dissertation at Technische Universität München, Munich, Germany.

[11] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013. ACM.

[12] F. Eyben, M. Wöllmer, and B. Schuller. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6, Sep. 2009.

[13] T. Gan, Y. Wong, B. Mandal, V. Chandrasekhar, and M. S. Kankanhalli. Multi-sensor Self-Quantification of Presentations. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 601–610, New York, NY, USA, 2015. ACM.

[14] German National Library of Science and Technology. Search as Learning - Investigating, Enhancing and Predicting Learning during Multimodal Web Search. https://www.tib.eu/en/research-development/project-overview/project-summary/salient, 2018. [Online; accessed 28-February-2019].

[15] F. Haider, L. Cerrato, N. Campbell, and S. Luz. Presentation Quality Assessment Using Acoustic Information and Hand Movements. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2812–2816, March 2016.

[16] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. RASTA-PLP speech analysis technique. In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 121–124, 04 1992.

[17] D. J. Hermes. Measurement of pitch by subharmonic summation. *IPO Annual Progress Report*, 21:24–33, 1986.

[18] R. Hincks. Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, 33(4):575 – 591, 2005.

[19] J. F. Kenney and E. S. Keeping. Root Mean Square. In *Mathematics of Statistics*, volume 1, pages 59–60. Princeton, NJ: Van Nostrand, 1962.

[20] J. Li, Y. Wong, and M. S. Kankanhalli. Multi-stream Deep Learning Framework for Automated Presentation Assessment. In *2016 IEEE International Symposium on Multimedia*, pages 222–225, Dec 2016.

[21] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[22] S. Poria, N. Howard, G.-B. Huang, A. Hussain, and E. Cambria. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2015.

[23] P. H. Ramsey. Critical Values for Spearman's Rank Order Correlation. *Journal of Educational Statistics*, 14(3):245–253, 1989.

[24] B. W. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. R. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 2013.

[25] R. Smith et al. Tesseract OCR [Computer program]. `https://github.com/tesseract-ocr/tesseract#tesseract-ocr`, Python Wrapper: `https://pypi.org/project/pytesseract/`, 2019. [Wrapper Version 0.2.6; retrieved 18-May-2019].

[26] J. P. Teixeira, C. Oliveira, and C. Lopes. Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology*, 9:1112 – 1122, 2013.

[27] B. L. Weathington, C. J. L. Cunningham, and D. J. Pittenger. *Understanding Business Research.* John Wiley & Sons, Inc., 2012. Dissertation at Technische Universität München, Munich, Germany.

[28] H. Yang and C. Meinel. Content Based Lecture Video Retrieval Using Speech and Video Text Information. *IEEE Transactions on Learning Technologies*, 7(2):142–154, April 2014.

[29] A. M. F. Yousef, M. A. Chatti, U. Schroeder, and M. Wosnitza. What drives a successful MOOC? An empirical examination of criteria to assure design quality of MOOCs. In *Proceedings - IEEE 14th International Conference on Advanced Learning Technologies*, pages 44–48, June 2014.

[30] E. Yumoto, W. J Gould, and T. Baer. Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, 71:1544, July 1982.

Acronyms

ASD Average Syllable Duration. 20

F_0 Fundamental Frequency. 9

HNR Harmonics-to-Noise Ratio. 12, 20, 29

PVQ Pitch Variant Quotient. 20

RMS Energy Root Mean Square Energy. 20

SIFT Scale Invariant Feature Transform. 17