A Framework for Automatic Topic Segmentation in Video Lectures

Eduardo R. Soares
Post Graduate Program in Computer Science
Federal University of Juiz de Fora
Juiz de Fora, Minas Gerais
eduardosoares@ice.ufjf.br

Eduardo Barrére
Post Graduate Program in Computer Science
Federal University of Juiz de Fora
Juiz de Fora, Minas Gerais
eduardo.barrere@ice.ufjf.br

ABSTRACT

Nowadays, video lectures are a very popular way to transmit knowledge, and because of that, there are many repositories with a large catalog of those videos on web. Despite all benefits that this high availability of video lectures brings, some problems also emerge from this scenario. One of these problems is that, it is very difficult find relevant content associate with those videos. Many times, students must to watch the entire video lecture to find the point of interest and, sometimes, these points are not found. For that reason, the proposal of this master's project is to investigate and propose a novel framework based on early fusion of low and high-level audio features enriched with external knowledge from open databases for automatic topic segmentation in video lectures. We have performed preliminary experiments in two sets of video lectures using the current state of our work. The obtained results were very satisfactory, which evidences the potential of our proposal.

KEYWORDS

Topic segmentation; Video lectures; Automatic Speech Recognition, Semantic annotation, Knowledge base, Content processing, Natural Language Processing

1 INTRODUCTION

Although videos have always been considered a powerful source to transmit knowledge, nowadays its importance is even more present in everyday life. That is mainly due the fact that advances in multimedia and communications provided the means to creation of interactive and robust online educational systems where video lectures are made available [24]. The academia, for example, has widely embraced e-learning model. Many universities offer distance learning courses aiming to reach students who do not have access to the campus. Just as companies which want to provide training to its employees and ensure that they learn new abilities without take too long for that.

Unluckily, the popularization of digital video has brought with it the difficulty of users find relevant content, on video repositories, according to their interests. This is due to the fact that there is an overload of information in video format on the web [9]. For this reason, one of the biggest challenges that information retrieval researchers face today is allowing users to access relevant information for their searches in the midst of so much content available. When we talk specifically about video lectures, it is very common that students watch the video to learn about just a few specific topics. But, typically, to quickly access this information is not an easy thing to do. Frequently, students spend a lot of time to localize interest points in a video. This occurs, mainly, because the unstructured and linear nature of video that does not provide navigability through contents that is ideal for learning [11].

Topic segmentation is the most common preprocessing to allow navigability through video lecture content. And it is consensual that this is able to turn information retrieval more agile. In addition, a topic-segmented video lecture can enhance learning across distance learning platforms by providing students with structured video content so they can navigate instantly from one topic to another whenever they choose, at their own learning pace [23, 24].

Despite all the advantages of segmenting video lectures into topics, that is not a trivial task. Human segmentation is very accurate, but the time required to accomplish that task manually it almost impractical, especially in large repositories of video lectures [11]. That is why many methods were proposed in the literature to automatically extract topic structure in video lectures through performing analysis in their features. Although these methods can achieve acceptable results, they ignore the fact that there is background information that cannot be obtained just with features of the video itself and that could be used to improve topic segmentation. For example, there are some external open knowledge bases like DBpedia [2] that can be used to enrich video features with this objective. Following this reasoning, we propose the use of low and high-level audio track features supported by external knowledge through an early fusion to obtain a topic segmentation for video lectures. Our main hypothesis is that, as our proposal uses different levels of audio track features added with external knowledge, it can obtain different relevant information from video lecture that combined contribute to a more complete understanding of the addressed topics in the video. And, that understanding may be used to obtain an accurate segmentation. Furthermore, when available, textual features from slides, textbooks, and metadata can also be used by our framework to improve topic segmentation.

In this work, we consider the automatic topic segmentation task as: given a video lecture V as input, determine the start and end time of each existing topic T_i in V. Where, we define a topic as a unit of a video lecture that is composed of sequential chunks that cover a same subject.

This paper is organized as follow. In section 2 we talk about the related researches of literature that deal with automatic topic

In: XVIII Workshop de Teses de Dissertações (WTD 2018), Salvador, Brasil. Anais do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres. Porto Alegre: Sociedade Brasileira de Computação, 2018.

ISBN 978-85-7669-435-9.

segmentation problem for video lectures. Section 3 describes our proposal and the ideas behind it. Section 4 describes the current state of our work and the preliminary results. Finally, in section 5 our conclusions are shown.

2 RELATED RESEARCH

Due to the relevance of the subject of this present work to the multimedia and information retrieval area, many different approaches to automatically extract topic structures in video lectures have been proposed over the years. These approaches, generally, may make use of different modalities of video information (eg. audio, video, text), in different semantic levels (eg. high, mid and low-level).

For example, in [4], the authors proposed a method for automatically obtaining video lecture summaries through topic segmentation. Their method is based on highlighting important sentences that were spoken in the video. To do this, they extracted low-level acoustic characteristics and used them to assign an "importance factor" to each of these sentences. The characteristics used by them had already proved to be good at detecting a change of subject in spoken discourses by identifying points where the speaker emphasized [1, 5]. Some of them are pitch, volume, duration of syllable sounds, and pause rate. Just like in [4], Togashi et al. [19] also made use of those same acoustic features, but combined with higher level linguistic features like cue words and phrases, word repetition, terms frequencies, and sentence locations. As results, the authors reported that the combination of those low and high-level features performed better than using them separately.

Speaking of methods that use high-level language structures to detect topic boundaries in video lectures, Lin et al. [11] proposed a whole linguistics-based one. In their approach, there are two sets of features which they aim to extract: content-based and discourse-based. The content-based set is composed of linguistic structures like noun phrases, verb classes, word stems, and others. Unlike it, the discourse-based set is composed of pronouns and cue phrases. According to the authors, the content-based structures are more related to the lexical and syntactical meaning of the body of content, while the discourse-based set has more to do with the neighborhood of hypothetical topic boundaries. After that feature extraction, a vector space is built using the weights of each feature in fixed size windows of transcript sentences. Then the similarity between each neighbor window is calculated and, the final topic segmentation is obtained using a similarity criterion.

When we work with video lectures, it is necessary to consider the existence of some specific characteristics that may be used to improve topic segmentation. One of them is that, frequently, video lectures follow the content of a textbook. Thus, Yamamoto et al. [22] presented a method for topic segmentation in video lectures through the association of audio transcription coming from automatic speech recognition (ASR) with topics keywords from the textbook summary. For this, they create a vector space where there are vectors that represent transcribed sentences and that represent textbook topics. Then the similarities between sentences vectors and textbook topics vectors are calculated. And, each sentence is associated with a single textbook topic using those similarities. Although, the assumption that the textbooks will always be available is a great disadvantage.

Another important characteristic of video lectures is the consensus that they, generally, do not present any significant visual changes like it is notorious in other kinds of video (e.g movies, news, cartoons, etc) [11]. Yet, there are some visual features in this kind of video that can be useful in automatic topic segmentation. That is why researches were carried out in order to explore those features, either individually or by combining with other sources of information. Like in [10], where the authors proposed a method based on image processing techniques to automatically extract handwriting from the blackboard and, thus, to identify the cutting points of a video lecture to obtain a topic segmentation. In [15], a combination of visual and text features was proposed. In this approach, three sources of information are used: video, slides, and subtitles. For each one, the method searches for transitions cues that indicate a topic change. Lastly, the transitions points obtained from the previous step are merged to form the final topic segmentation. As results, the authors reported that the combination of those transitions points improved significantly some evaluation ratios.

In [17], the focus is to automatically summarize lectures slides. The motivation behind that work is that it is very common for students to be asked to prepare for classes before they happen. The authors presented a method that can improve students preview through the use of visual and textual resources. To evaluate their approach in a real application, the authors conducted researches with more than 300 students. Their findings show that the use of summarized lectures slides by their method did not impact negatively on students performance and were capable of reducing their preview time. Slides can be a rich source of information in a lecture video. Furini et al. [7] published a recent work where they proposed the use of low-level audiovisual features combined with Optical Character Recognition (OCR) on slides to obtain a topic-based playlist, and thus to improve the information search on video lectures. Points of topic change are detected by low-level audiovisual features, and OCR is performed on slides to extract content information and allow keyword searches.

Still talking about multimodal approaches, Kishi and Goularte [9] proposed a method for automatic video scene segmentation where the features of different information channels are combined by computing their co-occurrences in shots before the segmentation step. This type of approach is called early fusion. Otherwise, if the approach takes individual decisions for each information channel and, after, those decisions are combined, it is called a late fusion approach. Although the Kishi and Goularte [9] method was proposed for automatic video scene segmentation, it provides a generic way to combine multiples sources of information that can be also used to segment video lectures into topics.

So far, we have only mentioned works that proposed methods which extract all information from video lecture itself. But, as we defend, there is background information that cannot be obtained just from the video. In this sense, there are researches of literature that have explored the use of external knowledge bases on topic segmentation task. In the work of Lin et al. [12], was presented an approach for automatic topic segmentation in video lectures that makes use of low-level audiovisual features, speech transcript, plus information from two types of lexical knowledge bases. The first one is a base of general words that are organized in synonyms

sets that are connected by semantic relationships, while the second is analogous to the first but specialized for lecture domain. More recently, a late fusion approach, that uses a knowledge base, was proposed by Shah et al. [16]. Their approach computes topic boundaries using subtitles, visual features, and Wikipedia articles, in a separate way. Then those topic boundaries are combined to generate the final set of topic boundaries. To find topic boundaries using subtitles and visual features, they used already cited methods from [11, 15]. To get topic boundaries using Wikipedia articles, they proposed a novel method that consists of segmenting into blocks the Wikipedia article that has the same subject as the video. And then, the method finds blocks of words from subtitles that most closely matches with Wikipedia article blocks to be the topic boundaries.

In this section, we briefly presented the state-of-the-art researches in automatic topic segmentation for video lectures. It was important to understand the tendencies of this area and situate the readers about the inspirations that we had when proposing our framework. In this paper, we propose a framework that makes use of early fusion of audio track low and high-level features combined with external knowledge. The advantage of our proposal is that it does not make assumptions about the existence of slides, textbooks or metadata associated with the video lecture, but it is capable to incorporate those information sources, when available, to improve the automatic topic segmentation.

3 PROPOSAL OF MASTER'S PROJECT

What we propose as master's project is a novel framework that is able to extract, enrich semantically and combine features from different sources, at different semantic levels (low and high), with the objective of segmenting video lectures into topics. Some of those features will not be essential for the operation of the framework, and will only be considered when their source is available for a specific video lecture. The semantic enrichment will be done through knowledge bases from where video lecture concepts and their relationships will be retrieved. In the end, an algorithm will take the segmentation decisions based on all extracted information. Figure 1 illustrates the proposed framework.

The main idea for obtaining this framework is to conduct an extensive study about techniques and methods of extraction, preprocessing, semantic enrichment and combination of video lecture features. Furthermore, the algorithms that use those features to segment the video lecture into topics will also be studied and analyzed. In the end, we expect to have a generic framework where it is possible to combine features from different sources related to a video lecture, when available, to obtain an accurate topic segmentation. The only mandatory source to the framework will be the audio track from which low and high-level features will be extracted since most of the information in a video lecture is in the teacher's speech.

3.1 Sources of information

Video lectures may have multiple sources of information, and thus, it is needed that the framework can handle that. The audio track will be the main source of information. Since the content of teacher's speech has a great importance on identifying the subjects of a video lecture. Thus, the audio track is expected to always be available.

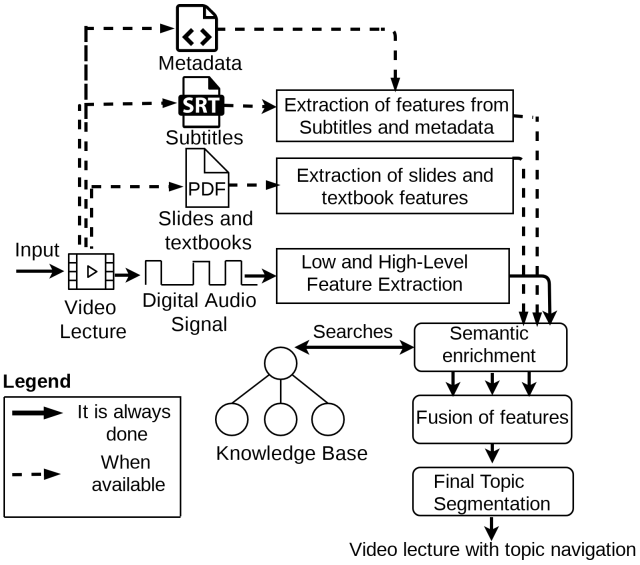

Figure 1: Proposed Framework

Other sources of information that will be considered by our framework, when available, are subtitles, slides, textbooks, and metadata such as LOM (Learning Object Metadata) [8].

Different from other literature approaches, our proposal is only dependent on the audio track. The other sources of information are not essential to framework operation but can be used when available to improve topic segmentation by bringing more information about video lecture's content.

3.2 Extraction of features

The extraction of features from information sources is a very important process of our framework. In this step, the framework extracts features that will be semantically enrichment and used by the topic segmentation algorithm.

For audio track, low and high-level features can be extracted. Some of the possible low-level features that will be considered by the framework are silence detection, fundamental frequencies (f0) and voice power density estimations. As stated in the previous section, these features have proven to be efficient in detecting moments of emphasis on speeches. And, these moments provide a good signal that there has been a change in subject. Since speakers tend to emphasize when they start a new topic. Another important low-level feature that we intend to analyze its contribution in topic segmentation task is the Mel-frequency cepstral coefficients (MFCC) [14]. Its use is widely diffused in the extraction of acoustic characteristics for automatic speech recognition. Therefore, MFCC is capable of giving a good acoustic representation along the audio track, which can be useful in topic segmentation task.

About the high-level features from the audio track, we intend to extract them in text format, through automatic speech recognition (ASR). The audio transcription is a powerful source of information because it makes possible to obtain the audio content in a format that allows the comparison between the contents of parts of the video, which can be used to identify a change of topics in the

video lecture. For the other sources of information, only high-level features will be extracted, also in text format. The textual content of metadata and subtitles can be extracted by parsing plain text files. For textbooks and slides, when the text content is rendered in PDF file, specific parsers can be used. But if the content is in Raster Matrix Format (RMF), both externally and within the video lecture, OCR (Optical Character Recognition) can be used for extraction [7].

All these textual content we could extract from video lecture may be submitted to Natural Language Processing (NLP) techniques like tokenization, stop-words removal, stemming, POS-tagging and others [3]. Furthermore, there are also well-known techniques to represent textual contents in a way that is possible to compute their similarities, like Bag-of-Words (BoW) and N-gram model [21].

3.3 Semantic enrichment

The semantic enrichment of features is a framework process where concepts found in feature extraction step will be searched in a knowledge base and, then, enriched with their relationships with the objective of adding context to them. The justification for this is that, without context, some concepts may seem to be unrelated, but when we analyze their relationships in a knowledge base we may find the opposite. And, that finding can improve topic segmentation decisions. For example, suppose that the concept "TCP" (Transmission Control Protocol) is extracted, through ASR, at the 30 seconds of a computer networks lecture. Later, at the 60 seconds of the same video lecture, the concept "UDP" (User Datagram Protocol) is extracted. Without any context, the topic segmentation algorithm may take the decision to separate those two video parts on different topics. But if we add some context to those concepts, we are able to discover that both TCP and UDP are Transport Layer Protocols and, therefore, are related. Thus, the segmentation algorithm may think it is better to change its previous decision and do not separate the video parts where those concepts were extracted.

To extract concepts from the features of the video lectures and enrich them, we intend to use the semantic annotation methods allied to searches in a knowledge base. To perform this, open and free knowledge bases like DBpedia[2] are available and can be used.

3.4 Topic segmentation

Choose an appropriate topic segmentation algorithm is an important issue to be solved in this master's project. There are two main classes of algorithms that can be used to perform that task: clustering and classification algorithms. The clustering algorithms will focus on try to find similarity and dissimilarity between parts of the video lecture, based on extracted features, to generate the video lecture topics. Well-known clustering algorithms that can be used are: *K-means*, *K-medoids*, *DBSCAN*, *Spectral Clustering*, and others [6, 20]. Instead, classification algorithms will use the extracted features to identify the existence or otherwise of a subject change in given segments of the video lectures. Classification algorithms that can be explored in this master's project include: *Decision Trees*, *Support Vector Machines*, *Neural Networks*, *Naive Bayes classifiers*, and others. In addition, because the performance of the algorithms can be influenced by the features used, as well as by the way we combine them, the evaluation of the techniques of early fusion of features is also part of our plans.

4 CURRENT STATE OF THE WORK

In this section, we will briefly discuss the current state of our work. First, we will present the current stages of processing performed by the framework. Then we will present the preliminary experiments that were performed and discuss the results obtained. A more detailed explanation can be found at [18].

4.1 Current framework processing

The current processing steps of our framework can be seen as a pipeline. In other words, the output of each stage of processing is the input of the next. The main idea of this approach is that, as the media flows through the pipeline, it is processed and transformed, allowing different semantic levels of information to be extracted from it. An overview of the proposed approach can be seen in Figure 2. From now on, we will use this figure as a reference to explain how the video lecture processing occurs.

To start the process, a video lecture is given as input and its audio track is extracted (i). Next, the extracted audio track is divided into chunks. That division is made so that the generated chunks do not contain silence, that is, after this stage, we have a sequence of fully voiced audio chunks (ii). After that, for each of these chunks, we extract a feature vector that relates the fundamental frequencies (f0) and power spectral density (PSD) [13] that occur over the chunk. Then, we use those vectors to compute and store an affinity matrix M_f that indicates the similarity between every pair of chunks according to those features (iii). In the next step (iv), those chunks are transcribed by an automatic speech recognition system. Thus, we obtain what was said in each chunk in text format. So in the fifth stage (v), we build a vector space with those transcriptions. In this space, each chunk s_i, where $i \in [0, 1, 2, ..., N-1]$ and N is the number of chunks, is represented by a vector and each dimension of this vector gives the weight of a word w_j in s_i. Where $j \in [0, 1, 2, ..., V-1]$ and V is the vocabulary size. This vocabulary can be predefined, or it can be built on the words transcribed in the video lecture for memory savings purpose. After that, another affinity matrix M_t is computed and stored. But this time, by considering the transcription features. The last semantic level of information extraction is done in steps (vi) and (vii). In the sixth step, each text from audio transcription is submitted to a semantic annotation. So after that, for each audio chunk s_i, we have a set of annotated terms A_i. Then, in the seventh step, for each set of annotated terms A_i, and for each annotated term $a_m \in A_i$, where $m \in [0, 1, 2, ..., G_i - 1]$ and G_i is the number of annotated terms in A_i, we search in a knowledge base for terms that have the same meaning as a_m. Next, we take those terms, including a_m, and we get their categories in the same base. After that, we have for each term $a_m \in A_i$ its synonyms and categories. So in the step (viii) is built a vector space and calculated an affinity matrix M_a, just like in step (v), but this time considering the features that were extracted in the steps (vi) and (vii). In the step (ix), we linearly combine all affinity matrices that were obtained in the previous steps to get a final affinity matrix H, used in the step (x) by the spectral clustering algorithm to generate the video lecture topic boundaries. These

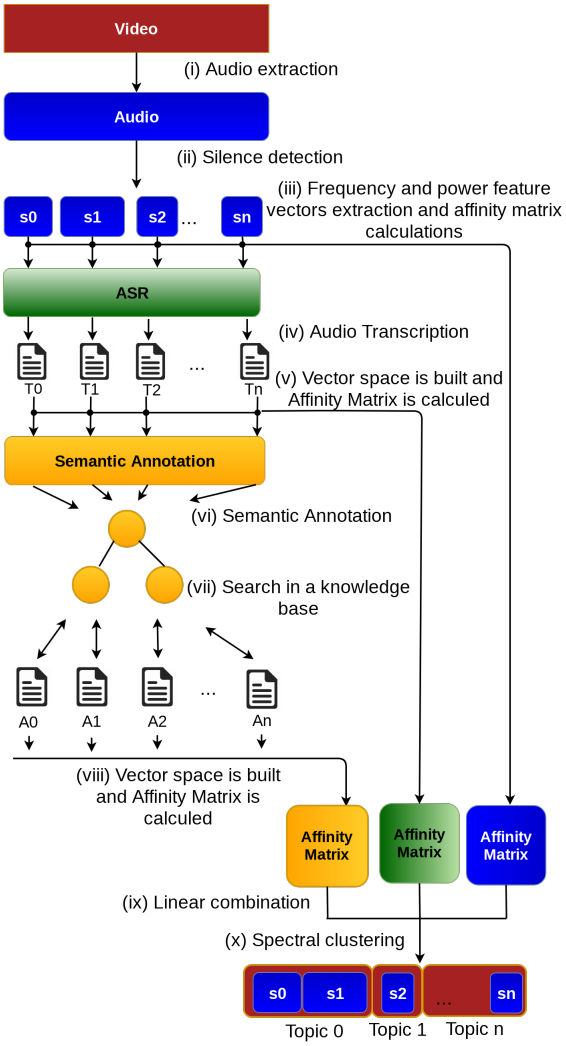

Figure 2: Overview of the current stage of the framework [18]

boundaries are obtained by clustering audio chunks according to H and, then, by using timestamps to map clusters of audio chunks into video lecture topics.

4.2 Preliminary Experiments

We conducted preliminary experiments with the objective of evaluating the impacts of incremental addition of higher-level features for topic segmentation. Thus, we separated the features of video lecture's audio track in three categories, from the lowest level to the higher: F, T and A. Where F are the frequency and power features, T represents the audio transcription features and A the features extracted through semantic annotation. Then, we consider that we have a sequence C of 3 sets formed by the combination of those categories in a way that, the set $i+1$ of C has the elements of the set i plus one element of the next category. That is $C = \{\{F\}, \{F,T\}, \{F,T,A\}\}$. Then, for each video lecture v_j from the data set, we run the framework processing 10 times considering each set of features. So, for each set of features, we calculate the average of evaluation ratios obtained for that video lecture through the 10 executions. With this experimentation, we want to show that combining lower level features with higher ones increases the quality of topic segmentation. The need to execute our method 10 times for each set of features in each video lecture is that the used clustering algorithm is based on K-means, which is a random start algorithm. So by running Q times and taking the average, we obtain a more significant evaluation of our method.

4.2.1 Evaluation data set. Our evaluation dataset is composed of 44 video lectures in Brazilian Portuguese, where 34 of them were extracted from the Videoaula@RNP repository[1]. The video lectures from this repository already have a topic segmentation which we used as the ground-truth for comparison. The other 10 video lectures were extracted from YouTube, and they did not have a previous segmentation. We had to make our own manual segmentation to use as ground-truth in this case. The decision to merge videos from those two sources into a dataset was based on the fact that the video lectures from Videoaula@RNP were made in a more traditional way by following a well-defined script, unlike the YouTube videos selected. The YouTube video lectures were made in a freer style and have a shorter duration than traditional video lectures. In this way, we also want to evaluate the impact of those differences on our method's performance.

Since in our research we did not find any evaluation data set in Portuguese for automatic topic segmentation task, we make our dataset publicly available on Google Drive[2]. For the final project of master's degree, we also intend to evaluate our framework in a data set of video lectures in English.

4.2.2 Evaluation ratios. To evaluate our method, we compare for each video lecture the automatic segmentation with the ground-truth. For these comparisons, we have chosen the mean of precision, recall e F-measure across the ground-truth topics. These are typical ratios used in literature to evaluate the results of information retrieval and multimedia processing tasks. Along the development of the master's project, other metrics that can be good for evaluating the framework's performance in the task of automatic topic segmentation, such as Overflow and Coverage [9], can also be used.

4.3 Preliminary Results

By doing the experiments that were previously described, we got the overall results which are shown in the Table 1 and 2. While Table 1 shows the results obtained in the data set of video lectures from Videoaula@RNP, Table 2 presents the obtained results on the YouTube videos. As can be seen, in overall, the incremental addition of feature levels has improved the automatic topic segmentation in 5%, if we compare the first set of features $\{F\}$ with the last one $\{F,T,A\}$. This behavior was noticed in the two data sets, which is a good sign that by considering different levels of information we can get a more accurate topic segmentation.

Also, in this preliminary experimentation, it has to be noted that the overall obtained results in the two data sets were significantly

[1] http://www.videoaula.rnp.br/portal/home
[2] https://goo.gl/UfG88k

Table 1: Method performance on Videoaula@RNP videos

Set of Features	$\{F\}$			$\{F,T\}$			$\{F,T,A\}$		
Lecture	Prec.	Rec.	Fm.	Prec.	Rec.	Fm.	Prec.	Rec.	Fm.
Overall	0,59	0,61	0,57	0,63	0,64	0,60	**0,64**	**0.65**	**0.62**

Table 2: Method performance on YouTube video lectures

Set of Features	$\{F\}$			$\{F,T\}$			$\{F,T,A\}$		
Lecture	Prec.	Rec.	Fm.	Prec.	Rec.	Fm.	Prec.	Rec.	Fm.
Overall	0.85	0.80	0.80	0.89	0.83	0.83	**0.90**	**0.85**	**0.85**

distinct. According to our investigations, the reason for that is because the duration of the video lectures and their number of ground-truth topics impact directly on the difficulty of segmenting them into topics. This occurs because of a characteristic of the problem, in which errors in the formation of a topic are propagated to the following topic, then the final propagated error is directly proportional to the number of topics in the video lecture. That error propagation occurs because of when the method wrongly clusters an audio chunk s_j in a topic t_i, the precision of t_i is affected because of that error, just like the recall of t_{i+1}, where s_j should belong to. Thus, the average F-measure of an obtained topic segmentation also decreases. And, in our data set, the video lectures from YouTube has, on average, fewer topics than those from Videoaula@RNP. Another conclusion drawn from our analysis in the evaluation data set is that the videos extracted from YouTube present, in general, a much higher audio quality in relation to those from Videoaula@RNP. And, as all stages of feature extraction of our framework, that were implemented so far, are sensitive to noise, it is natural that a low quality of recording leads to more errors than a high one.

5 CONCLUSION

In this paper, we presented a master's project proposal which consists of a framework for automatic topic segmentation for video lectures. In our proposal, different levels of features from the audio track of a video lecture can be extracted and combined with features from other sources related to it. Like slides, textbooks, and metadata. We conducted preliminary experiments over the current stage of work that served to evidence for the potential of combining audio features of different levels and semantically enriching them through a knowledge base.

The next steps of this master's work , until the limit of March of 2020, consist of: constant updating of the bibliographic review; study and implementation of feature extraction techniques to explore the other sources of information; analysis of pre-processing and noise removal techniques; implementation and evaluation of techniques for early fusion of features and final topic segmentation; and lastly, the writing and defense of dissertation, and the publication of the results in papers.

REFERENCES

[1] Barry Arons. 1994. Pitch-based emphasis detection for segmenting speech recordings. In *Third International Conference on Spoken Language Processing*.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[3] Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 31.

[4] Xiaoyin Che, Sheng Luo, Haojin Yang, and Christoph Meinel. 2016. Sentence-Level Automatic Lecture Highlighting Based on Acoustic Analysis. In *Computer and Information Technology (CIT), 2016 IEEE International Conference on*. IEEE, 328–334.

[5] Francine R Chen and Margaret Withgott. 1992. The use of emphasis to automatically summarize a spoken discourse. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, Vol. 1. IEEE, 229–232.

[6] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Sebti Foufou, and Abdelaziz Bouras. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing* 2, 3 (2014), 267–279.

[7] Marco Furini, Silvia Mirri, and Manuela Montangero. 2018. Topic-based playlist to improve video lecture accessibility. In *Consumer Communications & Networking Conference (CCNC), 2018 15th IEEE Annual*. IEEE, 1–5.

[8] Wayne Hodgins and Erik Duval. 2002. Draft standard for learning technology-Learning Object Metadata-ISO/IEC 11404. *IEEE P1484. 12.2/D1* (2002).

[9] Rodrigo Mitsuo Kishi and Rudinei Goularte. 2016. Video scene segmentation through an early fusion multimodal approach. *Anais do XXII Simpósio Brasileiro de Sistemas Multimídia e Web* 2 (2016).

[10] Greg C Lee, Fu-Hao Yeh, Ying-Ju Chen, and Tao-Ku Chang. 2017. Robust handwriting extraction and lecture video summarization. *Multimedia Tools and Applications* 76, 5 (2017), 7067–7085.

[11] Ming Lin, Michael Chau, Jinwei Cao, and Jay F Nunamaker Jr. 2005. Automated video segmentation for lecture videos: A linguistics-based approach. *International Journal of Technology and Human Interaction (IJTHI)* 1, 2 (2005), 27–45.

[12] Ming Lin, Christopher BR Diller, Nicole Forsgren, Yunchu Huang, and Jay F Nunamaker Jr. 2005. Segmenting lecture videos by topic: From manual to automated methods. *AMCIS 2005 Proceedings* (2005), 243.

[13] Rainer Martin. 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing* 9, 5 (2001), 504–512.

[14] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083* (2010).

[15] Rajiv Ratn Shah, Yi Yu, Anwar Dilawar Shaikh, Suhua Tang, and Roger Zimmermann. 2014. ATLAS: automatic temporal segmentation and annotation of lecture videos based on modelling transition time. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 209–212.

[16] Rajiv Ratn Shah, Yi Yu, Anwar Dilawar Shaikh, and Roger Zimmermann. 2015. TRACE: Linguistic-Based Approach for Automatic Lecture Video Segmentation Leveraging Wikipedia Texts. In *Multimedia (ISM), 2015 IEEE International Symposium on*. IEEE, 217–220.

[17] Atsushi Shimada, Fumiya Okubo, Chengjiu Yin, and Hiroaki Ogata. 2017. Automatic summarization of lecture slides for enhanced student preview-technical report and user study. *IEEE Transactions on Learning Technologies* (2017).

[18] Eduardo R Soares and Eduardo Barrére. 2018. Automatic Topic Segmentation for Video Lectures Using Low and High-Level Audio Features. In *Proceedings of the 24rd Brazillian Symposium on Multimedia and the Web*. ACM, 189–196.

[19] Shingo Togashi, Masaru Yamaguchi, and Seiichi Nakagawa. 2006. Summarization of spoken lectures based on linguistic surface and prosodic information. In *Spoken Language Technology Workshop, 2006. IEEE*. IEEE, 34–37.

[20] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing* 17, 4 (2007), 395–416.

[21] Hanna M Wallach. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 977–984.

[22] Natsuo Yamamoto, Jun Ogata, and Yasuo Ariki. 2003. Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition. In *Eighth European Conference on Speech Communication and Technology*.

[23] Haojin Yang and Christoph Meinel. 2014. Content based lecture video retrieval using speech and video text information. *IEEE Transactions on Learning Technologies* 7, 2 (2014), 142–154.

[24] Dongsong Zhang, Lina Zhou, Robert O Briggs, and Jay F Nunamaker Jr. 2006. Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & management* 43, 1 (2006), 15–27.