NANYANG TECHNOLOGICAL UNIVERSITY

BROADCAST NEWS NAVIGATION

XU SHIYANG

SCHOOL OF COMPUTER ENGINEERING
2013/2014

NANYANG TECHNOLOGICAL UNIVERSITY

CZ4079
BROADCAST NEWS NAVIGATION

Submitted in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Computer Science
of the Nanyang Technological University

by

XU SHIYANG

SCHOOL OF COMPUTER ENGINEERING
2013/2014

Abstract

With the rapid development in technology, digital news video has become increasingly popular due to the convenience it can bring to the viewer. However, the linear structure of the video often poses difficulty for the viewer in accessing particular news that is of his/her interest. To make the news video more accessible, one possible way is to segment the video into small parts that are classified by their individual genres. However, manual segmentation is costly, thus automatic news segmentation is highly desirable. Current automatic news segmentation methods cover two major areas: topic boundary and topic detection. These techniques rely on common features mainly visual, audio and texts. However, some of the methods such as LDA are not suitable for broadcast news as the amount of texts is too little for topic detection. In this thesis, we experiment with a hybrid-based method, which is suitable for broadcast news video. This approach uses shot change detection, speaker change detection and natural language processing such as noun phrase extraction to generate a well-defined topic boundary. To materialize these concepts, a web based broadcast news navigation (BNN) is used. BNN provides web based retrieval tools from the multimedia database which support user searching. As part of the testing phase, our experiment results indicate that the segmentation technique outperforms other state-of-art techniques while using the hybrid-based method.

Acknowledgments

This Final Year Project was made possible thanks to the cooperation and support of a number of people, who have helped me gain much more than what was expected. I am grateful and thankful to them.

First of all, I would like to express my deepest appreciation to Prof Chng Eng Siong, my Supervisor, for his valuable advice and assistance during the development of the project. Though, he was very busy with his own work, he always showed patience towards me when correcting my mistake. His dedication and perseverance towards research and education has been the primary motivation for me to work hard in this project.

I also wish to thank Steven Du Shi Jian for his valuable advice and suggestion throughout this project.

Finally, to Emerging Research Lab staffs, which helped me in my project and provide support for all my software issues.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

In recent years, rapid development in technology has led to increased accessibility of digital news video across various media channels (e.g. YouTube, Facebook, News sites). Reports had shown that 31% of the tablet users are willing to spend more time with digital news video daily than reading newspapers [24]. However, the linear and unstructured features of the news video can restrict viewers in accessing news that are of his/her interests. For instance, he/she has to look through the entire video or even play back and forth several times to access the right news in the video and it becomes very troublesome when the duration of the videos is long. Therefore, effective management of these videos are essential for searching and retrieving of a particular news.

As broadcast news is structured and has clearly defined topics (*e.g. politics, economics, sports, weather, technology, etc.*)[8], the proposed solution is to divide the videos into small, independent units and classify these units based on a general topic. By using the technique of segmentation, viewers can navigate the news video and zoom into a particular topic of interest with ease. In addition, data analyst can also discover trends of topics that viewers are more interested in and provide more of such news in the future broadcast. While segmentation provides convenience to viewers, performing the technique manually is time-consuming and requires a human to watch the entire video and understand the context before dividing it into different segments. To address this issue, this thesis focused on developing a segmentation technique that divides and classifies news videos automatically by detecting topic boundaries.

1.2 Objectives and Aims

Automatic segmentation of broadcast news is a challenging problem in the research community since there is no clear solution in creating perfect boundaries. Many well-known algorithms have been developed, but each of them has its limitation. In this thesis, the primary objective is to study on these existing algorithms and understand the problems and limitations of each of them. These algorithms are applied in a prototype system which allows user to browse and navigate broadcast news.
The aims of this project are:

1. **News segmentation**: The key problem is to separate and identify news boundaries within broadcast news. Different methodologies had been used to solve this problem, mainly using visual, audio and text features, but neither had produced satisfactory results.

2. **News topic detection**: categories news stories to topics. In this way, information retrieval is supported. Therefore, it allows users to identify topic of interest and select the relevant news from the system.

3. **Automatic and scalable system**: Users would like to choose a news programme, and allow the system to do the rest of the work. In this thesis, the programme will be fully automated from the point user selects the news to the generation of an output news video with different news stories. Overtime, more news video will be added to the system. To scale up, a database system is used to store this information.

4. **Usability**: As this is an application used by typical users who may not be tech-savvy. Different features such as searching and filtering (covered in Chapter 4) are added to enhance user-friendliness.

1.3 Scope

The scope of this thesis restricts to the broadcast news domain. All the related process and methods are developed and tested only on broadcast news data. However, it is possible to use the technique that was proposed in this thesis for other domains.

1.4 Report Organisation

This report is divided into six chapters and an overview of each chapter is as follows:

- Chapter 1 provides an introduction of the project and a summary of its motivation and the scope.
- Chapter 2 discuss on the related works of various researchers that were used in this thesis.
- Chapter 3 provides an overview of the concepts/ideas used in the design of the news navigation prototype as well as the system specification.
- Chapter 4 focus on the implementation of the system.
- Chapter 5 addresses the experiment results and evaluation.
- Chapter 6 concludes the report with future direction and lesson learnt.

Chapter 2

Literature Review

The existing news Segmentation is broadly divided into two areas [9], establishing topic boundaries and detecting relevant topics. For the former, features are drawn from three ways mainly visual, audio or texts. Topic detection on the other hand discovers topics of each segment by text analysis. The results of the news segmentation are stored in a database. A broadcast news navigation(BNN) is created to support retrieval of information from the database.

2.1 Visual

The video is made up of a sequence of many images or frames. These frames grouped together to form a shot. Shots with graphical or semantically similar contents, group together to create a scene [15].

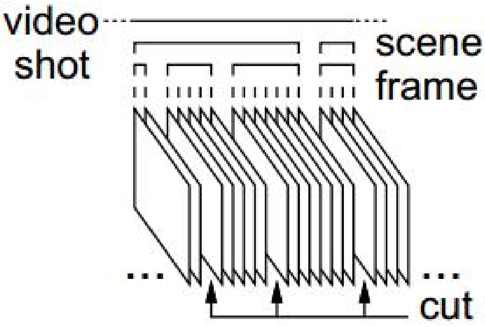


Figure 2.1: Video sequence

2.1.1 Shot Detection

When a cut occurred, the transition from one shot to another shot happens over two frames. These two frames are considered the key frames as they represent the start and end of two shots. In order to perform shot detection, various approaches exploit the colour, spatial and temporal dimension.

The most commonly methods rely on algorithms that use image cues like colour histogram which exploit the colour dimension. First, the colour percentages for a frame is stored, subsequently the results are compared with the next frame. If the difference is above a threshold, a shot change occurs. Colour histogram is easy to implement, but it only measure the distribution of colour, and was not able to measure the saturation in different area of an image. Therefore, shot change detection by these methods may not be accurate[4].

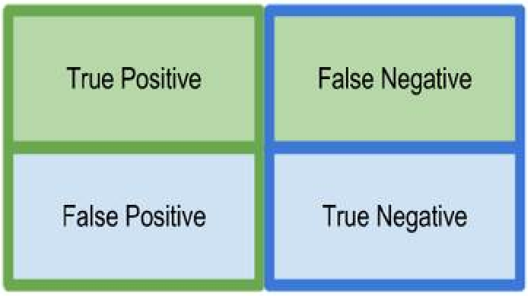
Shot change detection is often used in news segmentation as each shot may help in creating possible topic boundaries. Ide and Tanaka(2001) [22] introduced a method using shot detection. He done it by first segmenting the video into shots and used clustering techniques to classify each shot into graphically similar types. After that, caption detection is used to classify them into respective topic (*e.g. anchor, interview, etc.*). These classes required very structured domain in order to work properly.

2.1.2 Face Recognition

To overcome the limitation of shot detection, face recognition is often used in the news segmentation. Yu and Daneshi(2013) [6] considered the use of face recognition to detect the anchor. In news video, anchor is a person who presents and coordinates each news stories. The role of an anchor is significant in news segmentation as the multiple re-occurrences of the anchor helped establish topic boundaries [1][32]. The use of face recognition for detecting anchor is more reliable than shot detection since it is not sceptical towards global features. However, it is sensitive toward lighting conditions and face pose. Therefore it tend to create multiple clusters for the same person. To overcome this problem, Zhai and Yilmaz(2005) [44] proposed the use of "body" and face to define the anchor. It is based on the assumption that anchor wears the same dress throughout the entire news program. Thus, adding "body" provide better accuracy than just face recognition.

2.1.3 Metrics and Evaluation

Confusion matrix is a commonly used metric for evaluating shot detection:



True Positive	False Negative
False Positive	True Negative

Figure 2.2: Confusion Matrix

where:

- True Positive: actual *cut* is correctly classified
- False Positive: a frame is incorrectly classified as *cut*
- False Negative: actual *cut* is not detected
- True Negative: no frames are detected as *cut*

Different evaluation measures are used to measure performance of shot detections; precision and recall are two such measures that are commonly used by researcher [12], the former evaluating the number of false detections and the latter detect the number of missed detections:

$$\text{Precision} = \frac{\text{detects}}{\text{detects} + \text{false alarm}} \tag{2.1}$$

$$\text{Recall} = \frac{\text{detects}}{\text{detects} + \text{missed detects}} \tag{2.2}$$

2.2 Audio (speech)

Speech processing has been researched intensively in the past due to the rich source of information contain in it. The semantic interpretation of speech to text is not only useful knowing what was said, but also who say it. In a complex domain, such as broadcast news, speeches can come from an unknown number of speakers and there are seldom breaks in between utterances. Therefore, one of the approaches to identifying these speakers is through speaker change detection.

2.2.1 Speaker Change Detection

Speaker change detection(SCD) break up an audio stream into various homogeneous parts by recognizing the specific speech characteristics of speakers. Subsequently, it groups associated segment of speech coming from the same speaker in one cluster. One such technique that applied SCD is speaker diarization. It has the advantage of not requiring the data or number of speakers for training. It is done purely on unsupervised approaches that unlike some other speaker change detection methods.

2.2.2 Speaker Diarization

Speaker diarization is composed of four steps [27]. First, music and noise are removed. Next, voice activity detection (VAD) is carried out where speech is detected and segmented. It is followed by speaker change detection (SCD) where it detects speaker turn changes missed by VAD. After that, a hierarchical agglomerative clustering is used to group speaker segments from same speaker together.

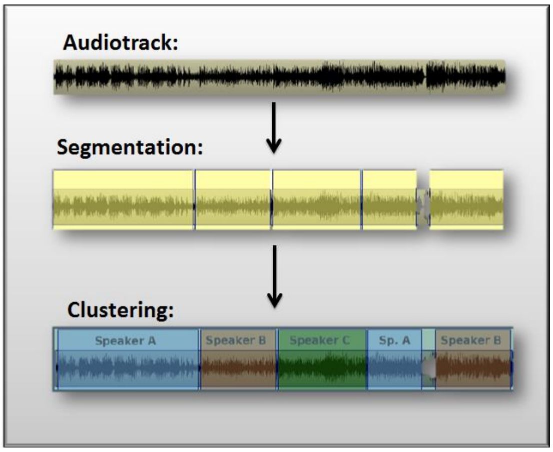


Figure 2.3: Diarisation Process

In broadcast news, speech are well-scripted and overlapping of speech rarely occurred [43]. Therefore, Speaker diarization is commonly used to analyse news segmentation. Huijbregts and leeuwen(2011) [30] proposed the used of statistics to measure the number of speech-line for each speaker. As most broadcast news only contain one anchor; the cluster that contains the most speech lines are classified as the anchor. The anchor then act as the boundary point. This classification technique is simple and effective in this restricted format. However, the uncertainty of the anchor's lines may pose complications in defining accurate boundaries.

2.3 Close Captions (text)

Topic segmentation and detection are categories that had been studies for years and there are still ongoing researches in this area. It applied many techniques used in information retrieval as well as natural language processing.

2.3.1 Topic Segmentation

In the process of segmenting a well define boundary. The first step is to do tokenization that breaks a stream of text into words. Subsequently, a list of stopword list is generate to remove token with no exact meaning (*e.g. a, the, is, them, they, etc.*). Each token is then sent for pre-processing using methods such as lemmatization and stemming. After that, tokens are stored in a table for comparison with subsequent tokens. This is based on the assumption that words that appear earlier in the document is most likely to appear again.

TextTiling

One such method is TextTiling, developed by Hearst(1997) [17] which automatically divides long texts into multiple paragraphs. It uses a sliding window method to compare the similarities between two neighbouring windows of text and subsequently, one window moves forward by an interval of two sentences and calculate the similarity again. This process is repeated until both the sliding windows reaches the end of the transcription. The places where the similarity between two windows is low are identified as potential topic boundaries. However, this method is often less reliable when used on non-technical paper such as news articles where repetition of vocabulary are minimum [14].

Similarity is determined by consine similarity:

$$\text{similarity} = \frac{\displaystyle\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\displaystyle\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\displaystyle\sum_{i=1}^{n}(B_i)^2}} \tag{2.3}$$

- if words are are the same in both boundary A and B, stories are the same.
- if there are no words in common, A and B are likely to be different stories.

2.3.2 Topic Detection

Two approaches for detection of topics are commonly used. For the first approach, topics are discovered based on clustering of on-line stream of text. Second approach used a dictionary of topics and identified the topic of the news story based on it.

POS Tagging and Chunking

Part of Speech Tag (POS) when combined with chunking is one such method that helps to dynamically generate topics through on-line stream of texts.
POS is a lexicon based library [42] that split the sentences into token with part-of speech-information. This tag classifies each of the word with their respective grammar (*e.g. Noun, Adjective, Verb, etc.*). Grammar that is considered interesting can then be extracted out for further processing. In broadcast news, stories usually consist of proper nouns words such as names of places, person or organisation. These noun words bear more semantic meaning than other part of the speech because they are the main characteristics used to identify news stories [26]. By using POS tagging, noun words were extracted out easily. However, noun words alone often does not provide a good understanding of the meaning [28].

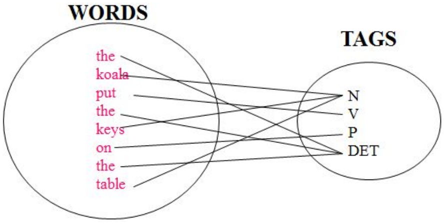


Figure 2.4: word-category information

In an attempt to make the noun more understandable, POS chunking is applied. POS chunking is the process of dividing sentences into non recursive inseparable phrases. Methods such as pattern matching and rule based methods help achieve these phrases. Some common noun phrase patterns are [28]:

- Preposition+Noun (*e.g. In the toilet, of his uncle*)
- Determiner+Noun (*e.g. As the cat, this dog*)
- Adjective+Noun (*e.g. Beautiful car, cheerful girl*)

In a method suggest by Panem and others(2013) [39], he combines the use POS chunking and URL extraction to mine sub-topics from a stream of text. He first extracts nouns and converts them to key phrases using POS chunking. Subsequently, key phrases are enhanced by finding the synonyms words contain in the URL of the tweet. The merit of this approach is the ability to come out with topics that are precise and little ambiguity. However, this approach relied on the use of external references, which may affect it reliability since references maybe changed over time.

Detection of Topic using Dictionary

A dictionary consisting of a list of topics is used to categorize each news story. For each news boundary, keywords in boundary are compared with topics in the dictionary. The boundary that matched a topic with the smallest distance is categorized under that topic. Although this method discovers topics easily, news stories belonging to one topic could be easily related to another topic. The problem is due to the same weight assigned to each word. Word such as "international" can be related to "Business" or "Politics" topics. Sma and Brun(2000) [5] suggested the use of TF-IDF to resolve this problem.

TF-IDF

In most document classification, TF-IDF(*Term frequency -inverse document frequency*)[16] is an important numerical statistic that is intended to reflect how important is a word is in a document of a dictionary. Sma and Brun proposed using different weight to assign each term to recognise how important it is in a topic. For example, word such as"Bank" carried more weight in "Business" topic compare to "Politics".

TF-IDF can be deduced based on following equation:

$$W_{t,d} = (1 + logtf_{t,d}) \times log_{10}(\frac{N}{df_t}) \tag{2.4}$$

where tf denotes the frequency of a term t that occurred in a topic d, N denotes the total number of word in the dictionary and df denotes the frequency of a term t that appear in all topics. Using this method, weight increase proportionally to the frequency of a term that appears in the topic but is also determined by its frequency across all topics in the dictionary. For instance, terms that occurred frequently in a topic may no necessary be more frequent in all topics.

Latent dirichlet allocation (LDA)

Another method that is commonly used in topic detection is Latent dirichlet allocation (LDA). It is a generative model that automatically discovers topics within the text. Suppose a predefined set of topics is chosen from the dictionary. Then subsequently, noun words are extracted from the documents. For each document, the distribution of topic proportions is done, and each word is assign to a topic randomly according to the size of the proportion. A repetitive process is carried out to reassign words that are lower than a predefined threshold. Iterating the process for a period of time, it eventually reaches a stable stage where words are tag to topics accurately [11]. Misra [31] suggested the use of LDA for detecting topic boundary within a text-based document. However, the problem with LDA is the huge amount of text it required in order to perform substantially well [21]. Broadcast news which has a low amount of text is probably not a good option to use LDA.

2.3.3 Text Summarization

Text summarization is the process of extracting important details from the source to produce a whole new summarized text presented to users. Text summarization is used in broadcast news mainly to provide a summary for each news stories. In this way, users can have a snapshot of the news story before deciding to continue to watch the full segment. One of the earliest summarization techniques is proposed by Luhn(1958) [20]. He suggested the frequency of a particular word provides a useful indicator of its significance. Luhn then complies this list of word with its frequency. Sentences that contain any words from the list are extracted. All sentences are ranked in order of their significant factor. Top ranking sentences are finally selected to form the summary of the text.

2.3.4 Text Translation

Text translation by definition is the conversion of the source-language text to another equivalent target language text. It is widely used in broadcast news to appeal wider group of audiences. Melero and Avramidis(2012) [10] proposed the use of multilingual subtitles for video where the subtitles of the video are first extracted using ASR. Subsequently, captions are translated into other language using a third party system “Word Magic” through online calls. The use of “Word Magic”, however, only restricted to English and Spanish conversion.

2.3.5 Word Error Rate (WER)

Topic segmentation and detection by texts has its advantages and disadvantages. The advantage of using this approach is the ability to use the large body of research conducted on document text classifications. However, it has the disadvantage of having high error rates. Issues such as misspellings and omissions of text are common in news transcriptions. Therefore, WER is used to evaluate the performance of the text generated by Automatic Speech Recognition (ASR).

WER is derived from levenshtein distance that is used to calculate the distance in words between the ASR word sequence and the reference word sequence[34]. The distance to be computed as:

$$\mathrm{WER} = \frac{\mathrm{S+D+I}}{\mathrm{N}} \tag{2.5}$$

where S denotes the number of substitutions, D denotes the number of deletions, I denotes the number of insertions and N denoted the number of words in the reference text.

A lower word error rate shows better accuracy in recognition of speech, compared with higher word error rate.

2.4 Hybrid Approach

Lastly a hybrid approach that combines the use of visual, audio and text is proposed by Merlino and others(1997) [2]. Merlino suggested the use of visual cues to detect anchor or logo, audio to detect period of silence which maybe the start and end of commercial boundaries and the use of text (close captions) to define possible topic boundary point.

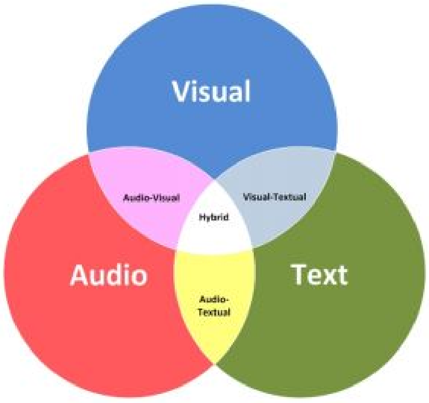


Figure 2.5: hybrid approach using low-level features

Hybrid approach has the advantage of covering the limitation of a particular feature. However, Jiang and Lin [19] discussed the difficulty of using a hybrid audio-visual approach. They stated that, for each set of feature, it has its own sequence of shots belonging to a particular scene. It is not clear which features carried more weight in making the final decision. Thus, this often led to poorly segmented boundaries.

2.5 Information Retrieval

The use of segmentation splits news video into many news stories. In addition, each news story may contain a summary, a topic or multiple key frames. This information is stored in a database. The web-based user interface is then implemented where users can search and browse videos by retrieving information from the database. This idea of a web based broadcast news navigation (BNN) is implemented by Merlino and others(1997). He introduces a search bar in BNN that enables users to enter a keyword and for each request, the database returns a list of relevant news stories. When users click on the news story, they can watch from the start of the segment to the end of the video with full VCR controls.

Chapter 3

Proposed Approach and System Specification

3.1 News Structure

Most news videos have a rather well-defined structured. It usually starts with intro/highlights that provide viewers sneak preview of upcoming news. Main body of the news contains a series of stories organised in broad categories such as politics, business, sport, entertainment, technologies [8]. Each news story normally begins with an anchor-person and continues to the reporter phrase and subsequently transit back to anchor to start a new story. In between the news, there are often a couple of commercial breaks. The structure manner of news videos provide opportunities to draft up story boundaries dynamically in an easy way.

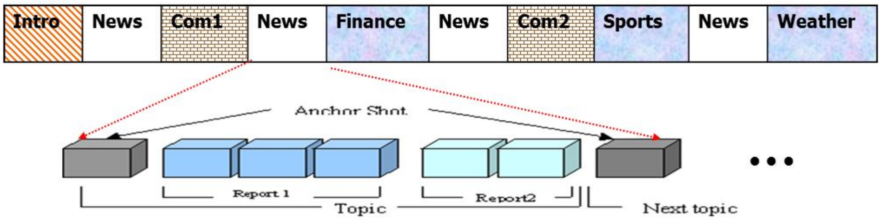


Figure 3.1: Broadcast news structure

3.2 Project Methodology

Among the approaches presented in chapter 2, most of the previous works has its pros and cons, and neither one could produce a perfect segmentation for news.

In this thesis, we proposed a hybrid-based method similar to the approaches used by Merlino and others(1997) [2] which considers the use of two features: visual and audio for defining the boundaries. The hybrid method helps to overcome each of its own weaknesses and works better for semantics modelling [23].Despite the problem stated by Jiang and Lin(2000) [19], a work-around solution is implemented (*to be explained in later chapters*).

In the visual approach, related works in chapter 2.1 had shown that the grouping of each shot to it respective class of topic may not be accurate and could affect the quality of topic boundary. Therefore in our proposed approach, in addition of shot detection, face recognition was used to detect the anchor.

In Audio approach, speaker diarization is used for detecting topic boundaries. By using a statistical method similar to Huijbregts and Leeuwen(2011) [30], an algorithm is used to identify the anchor. For each anchor, it defines potential boundary points.

In speech to text, ASR is used to extract textual information. However, due to the limitation of ASR; approaches such as TextTiling may not be feasible as the accuracy rate for news articles are poor[14]. Instead, texts are used for topic detection.

In our proposed method, the approach uses the method mentioned in chapter 2.3.2 to categorize boundaries to topics based on the best match comparison between each boundary and each topic in a dictionary. In addition to topic categorization, another approach applies a similar method proposed by Panem and others(2013) [39] which uses POS tagging and chunking to produce key phrases. These key phrases are used to define the news story headlines.

To address the concern for news indexing, BNN described in chapter 2.5 was built. All the multimedia data derived from the news segmentation were passed to the database. A search engine was provided in BNN to support retrieval of news stories from the database.

3.3 Definitions and Assumptions

3.3.1 Definitions

In this project, the following definition of each key term was made:

News

- News video: TV broadcast news that consists of one or more news stories.
- News story: A news story represents a complete, cohesive, news report on a particular topic.
- News segmentation: The process of doing topic segmentation and detection.

Visual

- Shot: It is the continuous footage or sequence between two cuts.
- Shot change: It is an abrupt change of frame content.
- Scene: It is a group of shots with graphical or semantically similar contents.
- keyframe: A frame that represents salient content of a shot.

Audio

- Speaker Change: A change in speaker contains in a conversation. In the news domain, *anchor* → *reporter* → *anchor* is a common pattern for speaker change.

Text

- Topic segmentation: The process of segmenting a news video into multiple news stories.
- Topic category: Structured defined topics that categorize each news story.
- Story Headlines: A title that describes the news story.

3.3.2 Assumptions

In this project, the following assumptions were made:

- Only English news videos are tested. Other languages are possible, but natural language tools for that particular language is essential for detecting topics accurately.
- Only one anchor is tested in this system. Detecting multiple anchors rely heavily on the transcription processed by the ASR, which is unreliable.

3.4 Agile Software Development

Agile approach was used in this project due to the dynamic environment where requirements may change. Agile approach promotes adaptive planning and encourages rapid identification to problems at stage with its iterative working style. The agile methodology developed fast where project scope is flexible and requirements may change. It also allows for delivering a small section of products at each iteration that provides an opportunity to test out small deliverables before integrating as a whole. With that, problems can be identified quickly as problematic areas are visible and can be solved more efficiently.
The project is separated into three different phases. They are research, development, and testing phases respectively. After completing these three phases, a feedback phase was conducted and different areas are reviewed. Subsequently, the three phases continue and an iteration cycle is formed. The iteration cycle stopped when a complete prototype is delivered and all the project objectives are achieved [41].

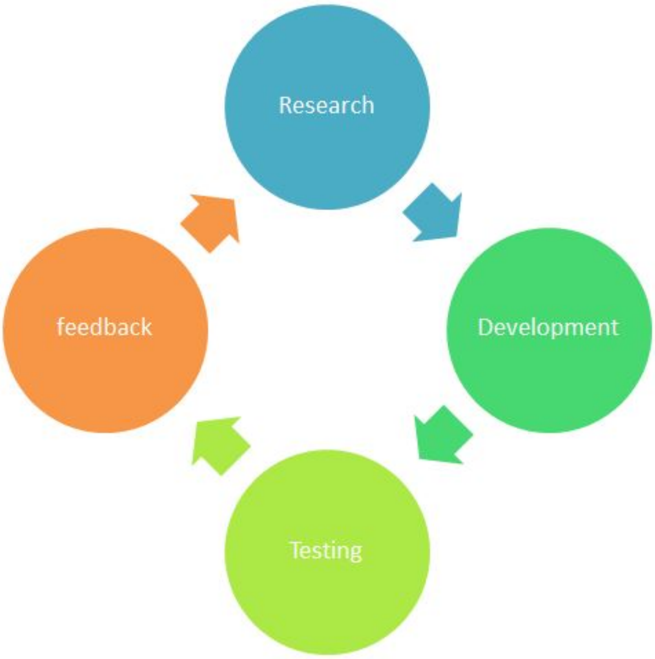


Figure 3.2: Agile approach used in this project

3.5 System Architecture

A pipe and filter model is used for the system architecture. The video source is first disassembled into frames and audio, where the latter is further pump into ASR to extract the text for text processing. After each individual part is completed, it was reassemble as a broadcast news navigation system.

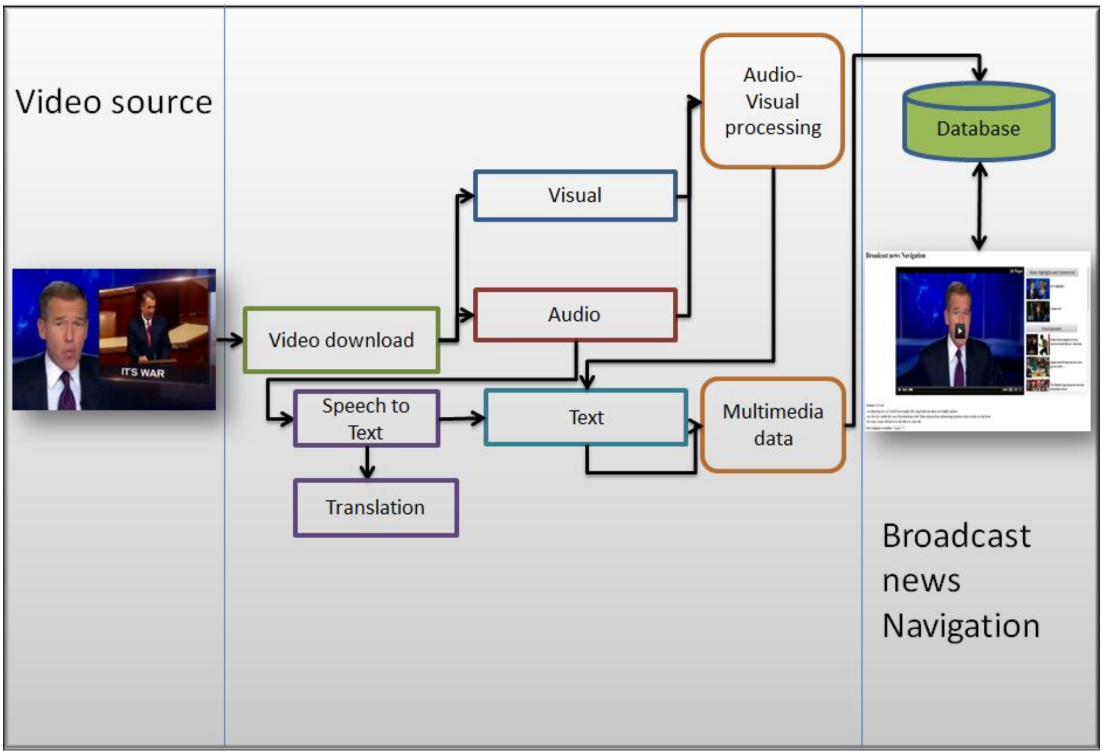


Figure 3.3: System Architecture

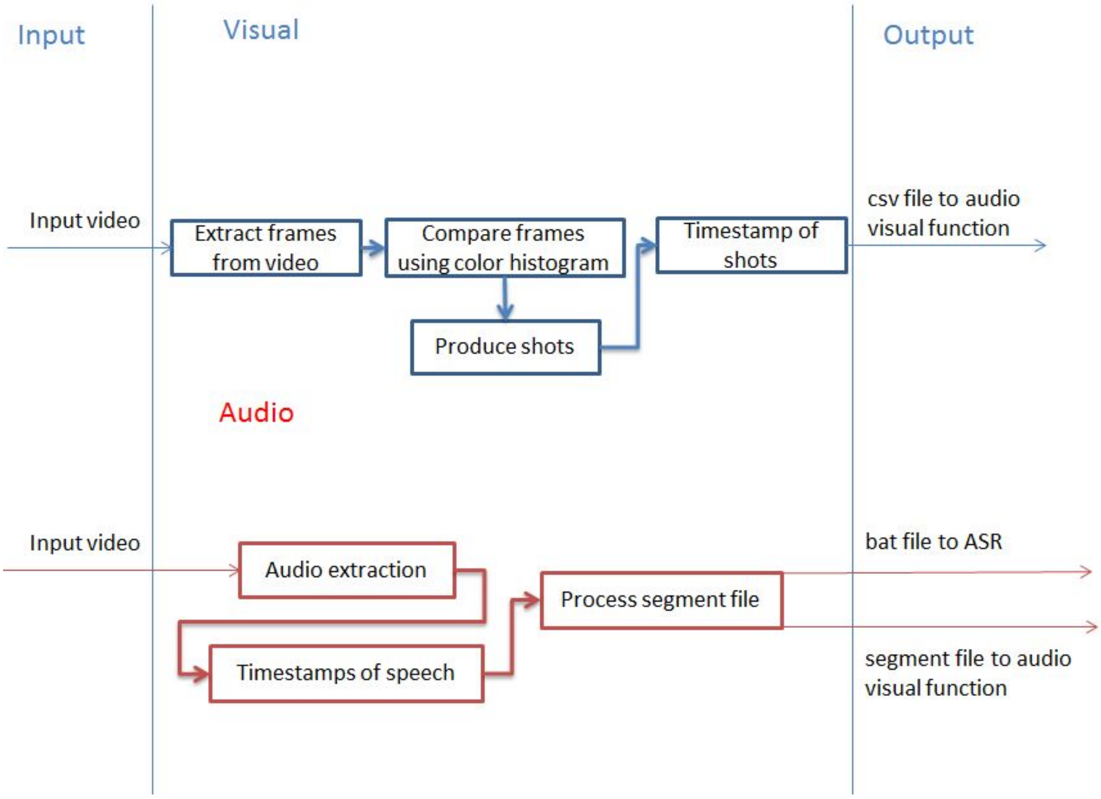


Figure 3.4: Visual and audio components

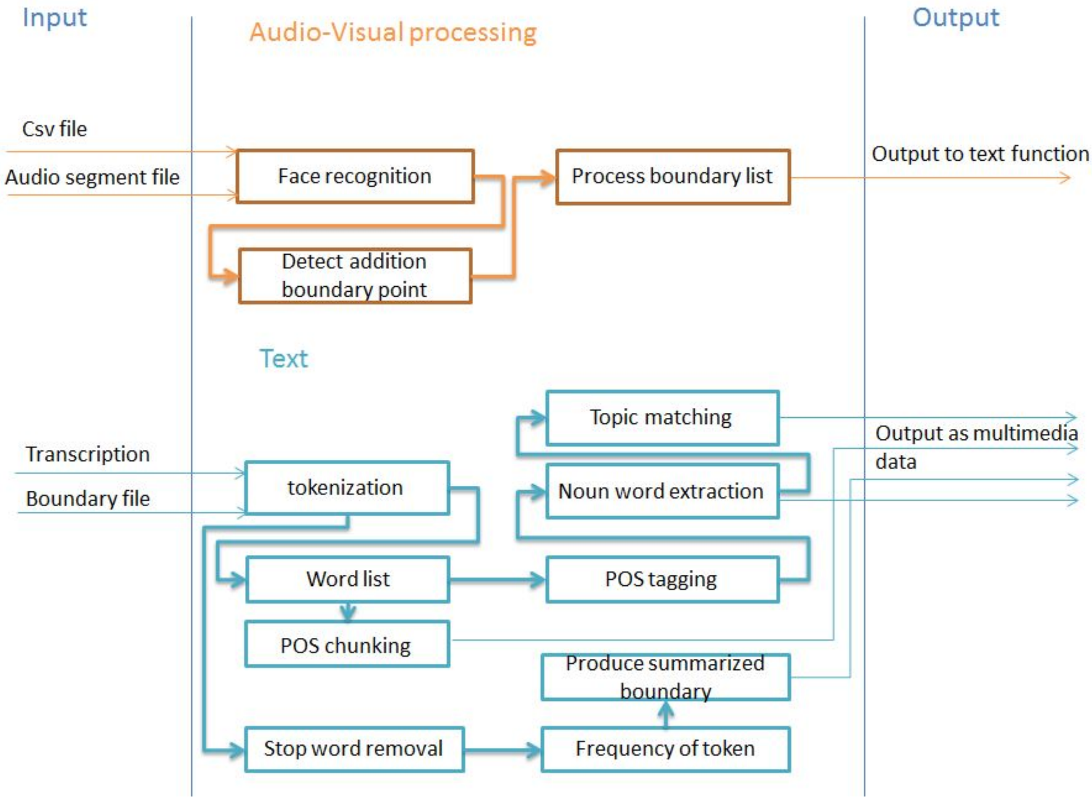


Figure 3.5: Topic boundary and text components

3.6 Component Specification

3.6.1 Video Download

- Videos of broadcast news are downloaded from YouTube using Python library called 'YoutubeDL'.
- Downloaded videos are converted into mp4 format and stored in a file directory.

3.6.2 Visual

- Video is segmented into individual frames using FFmpeg.
- Colour histogram is used to detect shots by finding the dissimilarity between frames.
- Timestamp for every shot is recorded in a csv file and passed to audio-visual function for further processing.

3.6.3 Audio

- Input video is sent for processing to extract the audio.
- Audio is segmented into a series of timestamps of speeches with each speaker change detected. The records are stored in a segment file.
- The segment file is processed to identify the *anchor* → *reporter* → *anchor* pattern. The rearrangement of speeches is updated in the segment file. In addition, the data also copy over to a bat file.
- The bat file invoked the execution of each audio segment for processing before sending to ASR for speech to text conversion.
- The segment file is sent to audio visual function for further processing.

3.6.4 Audio-Visual processing

- Both the audio segment file and the shot detection csv file are processed together.
- Face recognition is used for each shot having the timestamp within the duration of anchor speeches.

- The use of face recognition help to detect additional boundary points. The finalized boundary points are sent to text function for processing.

3.6.5 ASR

- ASR converts the speech to text for each audio segment based on the timestamps listed in the segment file.
- The transcription is then sent to text function for further processing.

3.6.6 Text

- The transcription from ASR is synchronized with the topic boundary.
- Texts in each topic boundary is split into individual tokens through a tokenizer and store in word list W and R.
- For each token in R, stop words are removed. The remaining tokens along with its frequency is used to extract significant sentences. These sentences were sorted in order and formed the summary.
- POS chunking is applied to extract noun key phrases from W and forms the story headlines.
- In word list W, POS tagging is applied to extract only noun words. These words are stored in a noun-list N.
- For each noun word in N, term frequencies are calculated and compared with the dictionary. Boundaries are classified into different categories based on scoring.
- In addition, all words from N are pass to the database for keyword retrieval.

3.6.7 Database

- All the multimedia data are passed to the database. Three tables are used mainly News, Stories, Keyword respectively.
- "News" table represents the collection of news video. All news video is located in a directory folder, while its name and id and pathname are stored inside the table.
- "News_Story" table includes news stories belong to a particular news video.

- “Keyword” table contains all the noun words for each news story of a news video.

3.6.8 BNN

- A web interface for BNN is created, with a video player embedded on the website.
- Interface contains a search engine for users to invoke a search.
- The results of the search return a list of information which allow users to find their desired news story quickly.

3.7 Summary of system workflow

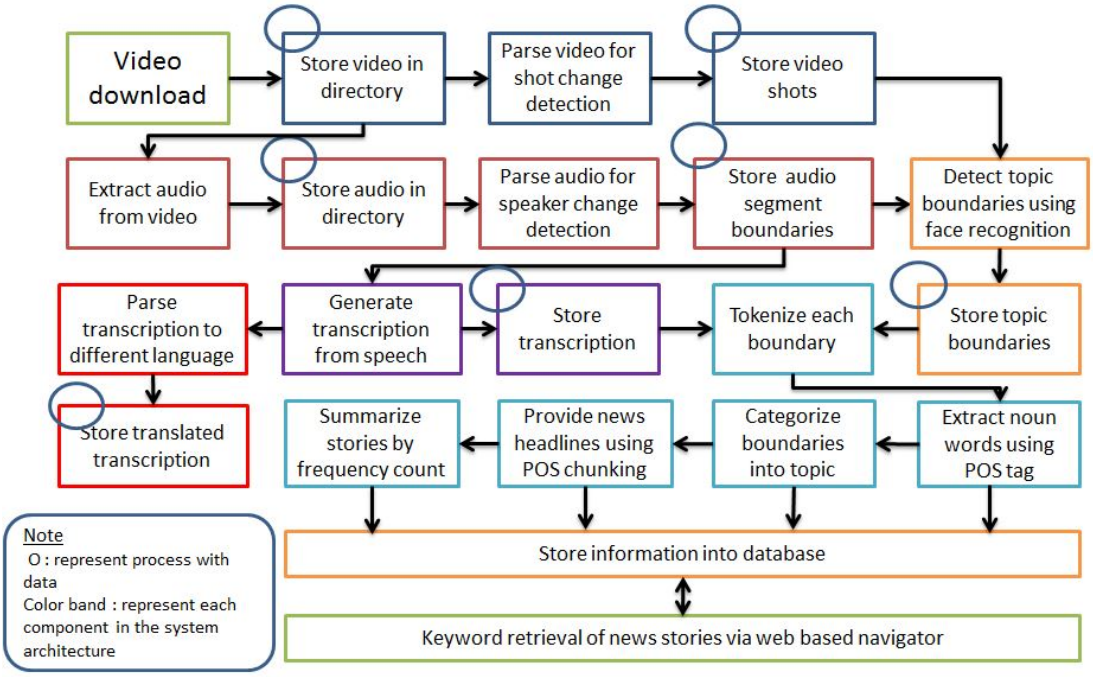

Video download
Store video in directory
Parse video for shot change detection
Store video shots
Extract audio from video
Store audio in directory
Parse audio for speaker change detection
Store audio segment boundaries
Detect topic boundaries using face recognition
Parse transcription to different language
Generate transcription from speech
Store transcription
Tokenize each boundary
Store topic boundaries
Store translated transcription
Summarize stories by frequency count
Provide news headlines using POS chunking
Categorize boundaries into topic
Extract noun words using POS tag
Store information into database
Keyword retrieval of news stories via web based navigator
Note
O : represent process with data
Color band : represent each component in the system architecture

Figure 3.6: System workflow

The figure above provides the summary of the entire system workflow. Data storing and passing are also explicitly stated in this workflow.

3.8 Tools and Technologies

The table below provides an overview of the environment setup of this project.

Environment setup

Hardware	4GB RAM, Intel core i7-3517 @1.90GHz
Operating System	Windows 7 Professional
IDE for development	Python tools for Microsoft Visual Studio 2012
Localhost server	XAMPP

Table 3.1: Environment setup

Languages and libraries

The extractions of different features (Visual, Audio and texts) were either obtained directly through third party applications (*e.g. Google Automatic Speech Recognition*) or built for our purposes (*e.g. topic boundary creation*). As a result, the news segmentation and BNN were created with a combination of different languages and libraries. The table below provides a summary of tools and technologies used in this project.

Component	**Language/Libraries**
YouTube video download	Python
Shot detection	FFmpeg
Face Recognition	OpenCV (C++, Python IDE)
Video Segmentation	Python
Audio extraction	VideoLan (VLC)
Audio processing	Sound eXchange (SoX)
Speaker change detection	Lium Diarization (Java)
Automatic speech recognition (ASR)	Google (C#)
Automatic text translation	Google translation toolkit
Text processing	NLTK (Python)
Web retrieval system	MySQL, HTML, JavaScript, PHP
Webplayer	JWplayer

Table 3.2: Languages and Libraries used

3.8.1 Python

Python is a programming language predominantly used for this project due to the existing libraries that are useful for building this application. For example, libraries such as Python image library (PIL) provide functions which convert images to a standard dimension, or to a greyscale image. These preprocessing steps done by the python library save a lot of our effort, which allow us to concentrate on the key components [36].

3.8.2 FFmpeg

FFmpeg is a command line program for transcoding multimedia files. It has a comprehensive library that consists of many functions that are suitable in segmenting video clips into numerous frames. In addition, timestamps of each frame are also generated. However, what separates it from other video processing software products is the capability of generating thumbnail size frames in less than a few seconds [3].

3.8.3 OpenCV

OpenCV is a computer vision library written in C/C++. It can be used via wrapper in several languages such as Python, Java, etc. OpenCV consists of a face recogniser class which is capable of detecting the faces of the people in the image. In this project, OpenCV was used in Python through a wrapper class, using the face recogniser class to identify the anchor in the series of images [35].

3.8.4 VideoLan

VideoLan is a complete software solution for many multimedia problems. It is capable of extracting audio from video file and convert the audio content to the desired format. The flexibility to choose the audio format and the numerous input options it supports is one of the reasons why it was used in this project. Besides, VLC allows the use of command line invocations, which support the extraction of audio without manual effort [18].

3.8.5 Sound eXchange

SoX is a utility tool that converts different formats of audio files into other formats. However, the main purpose of using SoX in this project is the ability to apply various modifications to the audio file such as the conversion of channels from stereo to mono. Besides, the script for splitting the audio file to multiple audio segments is also easy to implement [40].

3.8.6 Lium Diarization

Lium speaker diarization is an open source software that provides a complete toolkit for speaker diarization going from the audio signal to speaker clustering based on the CLR/NCLR metrics. These tools include MFCC computation, speech/nonspeech detection, and speaker diarization methods. The use of lium diarization produces a list of speakers and timestamps of their speeches, therefore, enabling the use of statistical methods to determine the anchor role [38].

3.8.7 Google ASR

Google ASR is a cloud based service in which a user submits audio data using a HTML Post request and receives a reply by the ASR in a form of a list.

One of the limitations of Google ASR was that audio data is limited to only ten seconds; longer clips are rejected and return no results. Besides, only "flac" format is accepted and the bit rate is at 16000. The user must also specify the language the audio file contains before the system could work.

Other issues such as customization of the system are prohibited. The entire ASR is treated as a blackbox, where the internal parts are unknown to users. In this restricted format, Google may update or change its service model, which may lead to changes in the returned data [13].

3.8.8 Google Translation

Google Translator toolkit is a web application that allows users to translate texts using the "Google Translate". Users can then upload and translate different articles such as Microsoft words, HTML text and Wikipedia articles. In this project, four kinds of languages are used: English, Melayu, Chinese and Tamil [25].

3.8.9 NLTK

NLTK provides a suite of libraries and programs for symbolic and statistical natural language processing (NLP) used in Python programming language. It provides libraries that are able to work on tasks such as tokenization, lemmatization, POS tagging and chunking. In this project, POS tagging is used predominately to extract noun keywords for matching with dictionary and POS chunking to extract noun key phrases [37].

3.8.10 MySQL

In this project, MySQL predominately used to storing all the multimedia information. This allows users to search through the database to locate a particular keyword in the news story. In addition, MySQL provides good support for web-applications and is open source. However, in the future, when the project scales up, the use of other proprietary databases such as Oracle or Microsoft DB will be preferred [7].

3.8.11 JWPlayer

JWPlayer is a Flash embeddable video player for websites. It supports a wide array of platforms and media formats. It is easy to configure, customize and extend. JavaScript is predominately used in working with JWPlayer functions. This provides the flexibility in working Javascript functions with JWplayer functions, and that is also one of the reasons JWplayer is chosen in this project [29].

Chapter 4

Implementation

In this section, a detailed illustration for each component is elaborated.

4.1 Video Download

The video was downloaded using a python external library called "YouTubeDl". This allowed the video to be downloaded from YouTube website using commandline script.

NBC Nightly News - 12/11/13 Full Broadcast

AirCastTV · 86 videos

204 views

Published on Jan 4, 2014
No description available.

Figure 4.1: Video download from YouTube

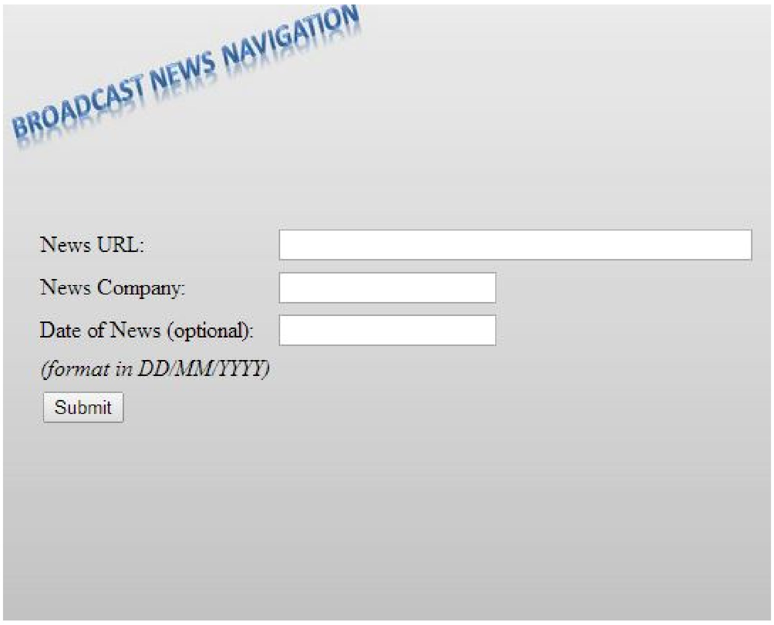


Figure 4.2: Web interface to download Broadcast videos

The Web interface allowed users to type in a particular broadcast news link. In this current set-up, we assume users will type in the appropriate video. Users are also supposed to type in the name and an optional date for naming purposes.

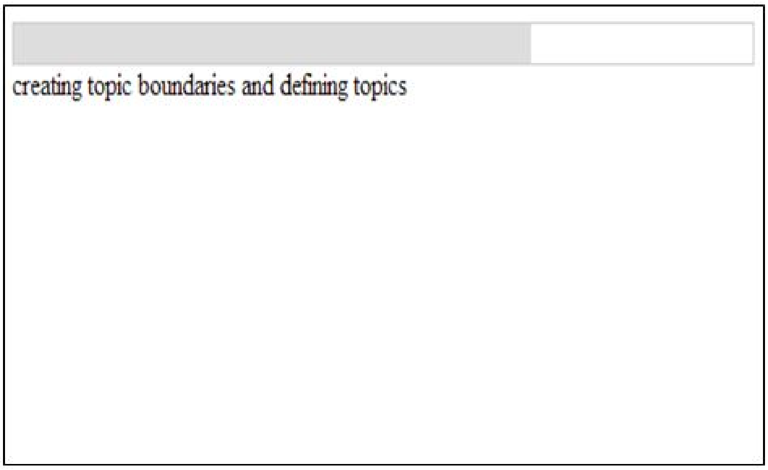

creating topic boundaries and defining topics

Figure 4.3: Progress bar used in the web application

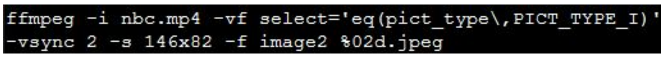
Once the user clicks the submit button, a progress bar is shown on the screen. It reports the current status of the system and informs users of how close it is to completing the set of tasks.

4.2 Visual

A good source of information in broadcast news program is the anchor shot. The anchor, which always reappears in subsequent intervals of the news program, may likely denote a news boundary. The first step to identifying the anchor shot in the news program is to divide the news video into a series of frames. This step is performed using FFmpeg software.

4.2.1 FFmpeg

To split the video into frames, the following script was used:

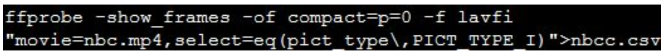


```
ffmpeg -i nbc.mp4 -vf select='eq(pict_type\,PICT_TYPE_I)'
-vsync 2 -s 146x82 -f image2 %02d.jpeg
```

Figure 4.4: FFmpeg script

- **vsync 2**: prevent FFmpeg to create more than one copy for each frame.
- **f image2**: write the frames into image files. Each image file can be dynamically created using a pattern. The pattern may contain %d that denotes a numeric value.

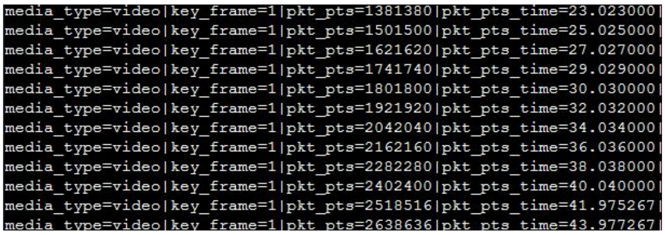
In order to generate the time stamp of each image, FFprobe (command-tool that shows media information) was used.



```
ffprobe -show_frames -of compact=p=0 -f lavfi
"movie=nbc.mp4,select=eq(pict_type\,PICT_TYPE_I)">nbcc.csv
```

Figure 4.5: A FFprobe script

- **show_frames**: provide information for each single frame in the input multimedia stream. Information in each frame consists of time in, time out and the duration.
- **of compact=p=0**: prints the information in csv format with p=0 changes the default value of the first line to 0.
- **f lavfi**: force FFprobe to use a special library called libavfilter. This library was able to perform certain condition in the video. In this case, select=eq(pict_type PICT_TYPE)



```
media_type=video|key_frame=1|pkt_pts=1381380|pkt_pts_time=23.023000|
media_type=video|key_frame=1|pkt_pts=1501500|pkt_pts_time=25.025000|
media_type=video|key_frame=1|pkt_pts=1621620|pkt_pts_time=27.027000|
media_type=video|key_frame=1|pkt_pts=1741740|pkt_pts_time=29.029000|
media_type=video|key_frame=1|pkt_pts=1801800|pkt_pts_time=30.030000|
media_type=video|key_frame=1|pkt_pts=1921920|pkt_pts_time=32.032000|
media_type=video|key_frame=1|pkt_pts=2042040|pkt_pts_time=34.034000|
media_type=video|key_frame=1|pkt_pts=2162160|pkt_pts_time=36.036000|
media_type=video|key_frame=1|pkt_pts=2282280|pkt_pts_time=38.038000|
media_type=video|key_frame=1|pkt_pts=2402400|pkt_pts_time=40.040000|
media_type=video|key_frame=1|pkt_pts=2518516|pkt_pts_time=41.975267|
media_type=video|key_frame=1|pkt_pts=2638636|pkt_pts_time=43.977267|
```

Figure 4.6: CSV containing media information

The result for performing FFmpeg produced a series of frames and its corresponding timestamp. However, these frames are not meaningful since each frame is highly correlated to a few frames preceding it with slight variation. Therefore, colour histogram is used. For each frame, the variance and the intensity are calculated over the luminance component and is compared to the next frame. If the value exceeds a given threshold T, a shot change occurred.

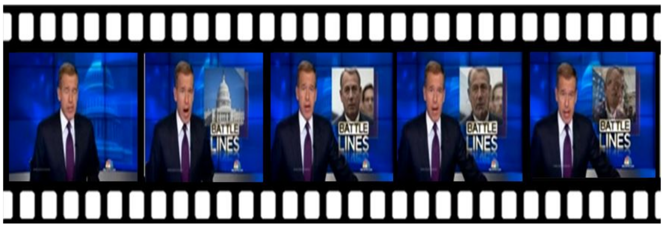


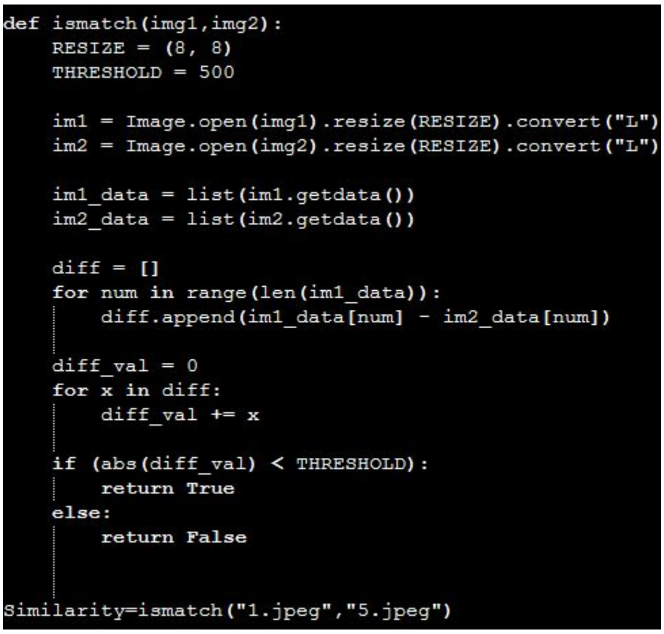
Figure 4.7: images with slight variation

4.2.2 Python Imaging Library (PIL)

In this project, Python Imaging library is used for our problem as it provides powerful image processing capabilities.

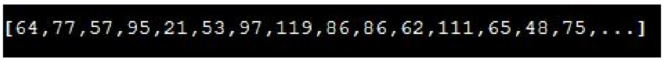
The python script which used PIL for colour comparison between two images is shown in the figure below.

- **Resize()**: normalize the image to a standard size.
- **Convert()**: convert the image to a grayscale image
- **Im.getdata()**: retrieve the RGB values of the pixels in the images



```
def ismatch(img1,img2):
    RESIZE = (8, 8)
    THRESHOLD = 500

    im1 = Image.open(img1).resize(RESIZE).convert("L")
    im2 = Image.open(img2).resize(RESIZE).convert("L")

    im1_data = list(im1.getdata())
    im2_data = list(im2.getdata())

    diff = []
    for num in range(len(im1_data)):
        diff.append(im1_data[num] - im2_data[num])

    diff_val = 0
    for x in diff:
        diff_val += x

    if (abs(diff_val) < THRESHOLD):
        return True
    else:
        return False


Similarity=ismatch("1.jpeg","5.jpeg")
```

Figure 4.8: python script using PIL



```
[64,77,57,95,21,53,97,119,86,86,62,111,65,48,75,...]
```

Figure 4.9: list of RGB values

Diff[] append the difference in rgb value for all pixels in the images. Diff_val was then compared with threshold T where T is a constant value. The result return true if the value is lower than T.

4.2.3 Limitations of Shot Detection

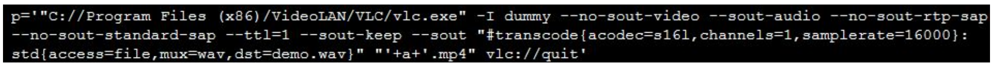
This process of shot change detection is completed, but it could not be used directly to detect topic boundaries as it will induce a significant number of missed transitions. This is because if the studio setting changes, the global features of the anchor shots will pose less similarity. Besides, there is also a possibility that colour histogram is similar for two shots of two different shots' content which happen to share the same colour information.

Therefore, the information of each shot are passed to the audio-visual function for further processing.

4.3 Audio

4.3.1 VideoLan

Audio is first extracted from the video via a VLC script which invoked the VLC cmd. The channel and sample-rate are all configured to 1 and $16000kHz$ respectively. The audio is also changed to *wav* format. This was done as Lium Speaker diarization only supports this configuration.

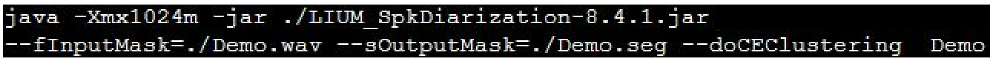


```
p='"C://Program Files (x86)/VideoLAN/VLC/vlc.exe" -I dummy --no-sout-video --sout-audio --no-sout-rtp-sap
--no-sout-standard-sap --ttl=1 --sout-keep --sout "#transcode{acodec=s16l,channels=1,samplerate=16000}:
std{access=file,mux=wav,dst=demo.wav}" "'+a+'.mp4" vlc://quit'
```

Figure 4.10: Script to perform audio capture

4.3.2 Lium speaker Diarization

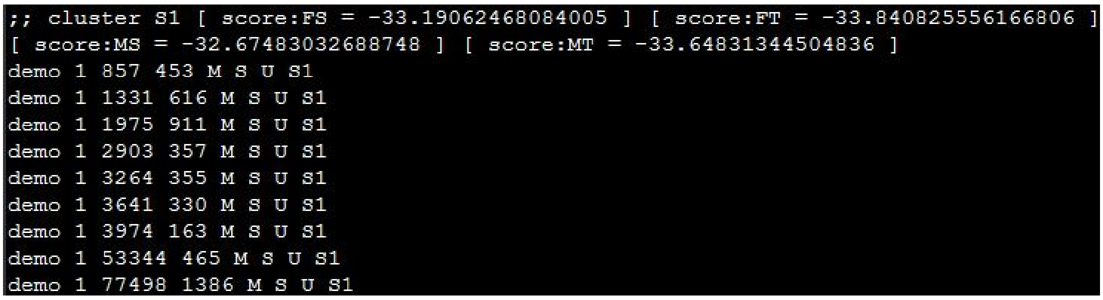
At the beginning of processing the audio, the identity of each speaker is not known and even the number of speaker is unknown. To detect speaker change and the number of speech for each speaker, Lium Speaker diarization is used. The following script to use Lium Speaker diarization is shown in the figure below:



```
java -Xmx1024m -jar ./LIUM_SpkDiarization-8.4.1.jar
--fInputMask=./Demo.wav --sOutputMask=./Demo.seg --doCEClustering  Demo
```

Figure 4.11: Script to perform speaker diarization

- **Xmx2024m**: set the memory to 2048MB, which is sufficient for an hour show
- **jar ./lium_spkDiarization.jar**: specifies the jar to be used.
- **fInputMask=./Demo.wav**: specifies the audio file to be used. The audio file must be in 16kHz and PCM mono. The use of mono ensures a single channel is used.
- **OutputMask= / Demo.seg**: is the output file containing the segmentation
- **doCEClustering**: computes the program using the CE Clustering. (CE=Cross Entropy), If other clustering are set, the program will be computed using the clustering.
- **Demo**: is the name of the show



```
;; cluster S1 [ score:FS = -33.19062468084005 ] [ score:FT = -33.840825556166806 ]
[ score:MS = -32.67483032688748 ] [ score:MT = -33.64831344504836 ]
demo 1 857 453 M S U S1
demo 1 1331 616 M S U S1
demo 1 1975 911 M S U S1
demo 1 2903 357 M S U S1
demo 1 3264 355 M S U S1
demo 1 3641 330 M S U S1
demo 1 3974 163 M S U S1
demo 1 53344 465 M S U S1
demo 1 77498 1386 M S U S1
```

Figure 4.12: Output of the Speaker diarization

The results of the diarization process produce the segment file. An example of the segment is shown in the figure:

- **Cluster S1**: contains all the speeches by speaker S1 in the entire clip
- **Score FS/MS**: represent the score for female/male wide bandwidth. The bandwidth score is computed based on comparing against the ESTER corpus.
- **Score FT/MT**: represent the score for female/male telephone. The telephone score is computed based on comparing against the French corpus that contains telephonic dialogues.

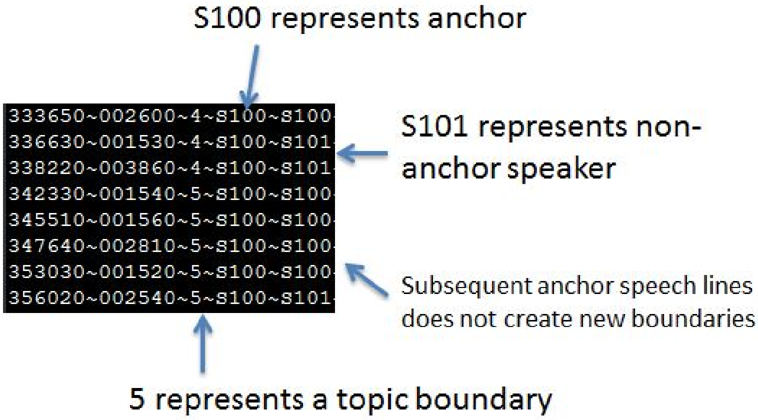
Start time and duration of the speech spoken by the speaker are explicitly stated in the cluster. To determine if the speaker is a male of the female, the FT/MT and FS/MS score are compared. The lower score determines the gender.

4.3.3 Identify Anchor and Creating Boundaries

In order to identify the speaker, one of the methods is to count the number of speech line in each cluster. In broadcast news video, the speaker that speaks the most is assumed to be the anchor and, therefore, in the segment file, a cluster with the most number of speech lines is classified as the anchor.

After the anchor has been identified, the speech lines are reorganised according to the timestamp. The topic boundary is created if and only if

- For each speech line belonging to the anchor speaker, subsequent speech lines were belonged to non-anchor speakers or,
- For each speech line belonging to non-anchor speakers, subsequent speech line were belonged to the anchor.



S100 represents anchor

```
333650~002600~4~S100~S100
336630~001530~4~S100~S101
338220~003860~4~S100~S101
342330~001540~5~S100~S100
345510~001560~5~S100~S100
347640~002810~5~S100~S100
353030~001520~5~S100~S100
356020~002540~5~S100~S101
```

S101 represents non-anchor speaker

Subsequent anchor speech lines does not create new boundaries

5 represents a topic boundary

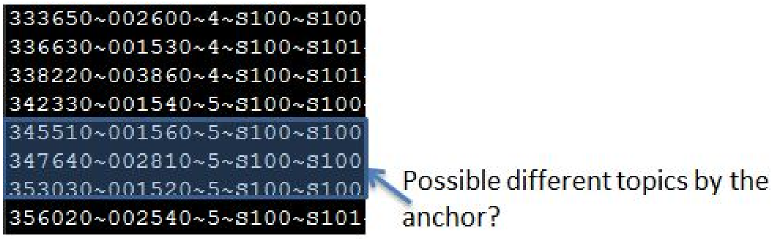
Figure 4.13: Boundary creation based on identifying anchor and non-anchor speech lines

4.3.4 Further Processing of Audio Segment

For each speech line, the duration is checked to ensure it does not exceed ten seconds. This was a limitation imposed by google ASR that only permit audio files that are less than ten seconds. Therefore, the solution is to split speech lines that timestamps are over ten seconds into two separate speech lines.

4.3.5 Limitation of Speaker Change

Speaker change method is used predominantly to detect a pattern e.g. (*anchor* → *reporter* → *anchor*) for defining topic boundaries. However, when the anchor decides to report news stories by him/her, the speaker change method would fail to recognise different topic boundaries. To overcome this limitation, the segment file is passed to audio-visual function, where further processing was done.



```
333650~002600~4~S100~S100
336630~001530~4~S100~S101
338220~003860~4~S100~S101
342330~001540~5~S100~S100
345510~001560~5~S100~S100
347640~002810~5~S100~S100
353030~001520~5~S100~S100
356020~002540~5~S100~S101
```

Possible different topics by the anchor?

Figure 4.14: limitation of speaker change

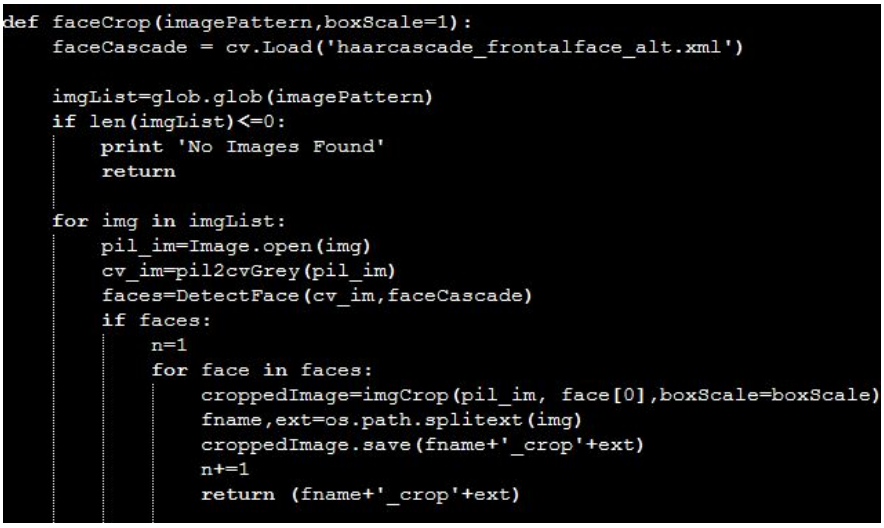
4.4 Audio-Visual processing

The use of just audio features has its limitations of not detecting true boundaries, and the use of just visual features may produce a high number of false positive boundaries. Therefore, shot details and audio segments are further processed in this function. To produce more accurate detection, face recognition is used. The entire process to extract faces from shots is done using OpenCV.

4.4.1 Face Recognition

In this thesis, haar cascade was used as the classification technique to detect the faces in the image. Haar cascade worked in a way such that in the detection phase, a window size is moved over to the input image, and for each subsection of the image the haar-like feature is calculated. The result is then compared to a learned threshold that separate non-objects from objects. However, haar cascade is a weak classifier and, therefore, require a large number of training in order to describe an object with sufficient accuracy. A good training run can take two to six days to get back a cascade. Therefore, in this project, an existing haar cascade created by Seo(2008) [33] was used.

A fragment of the python script which used haar cascade for face detection is shown in the figure below.

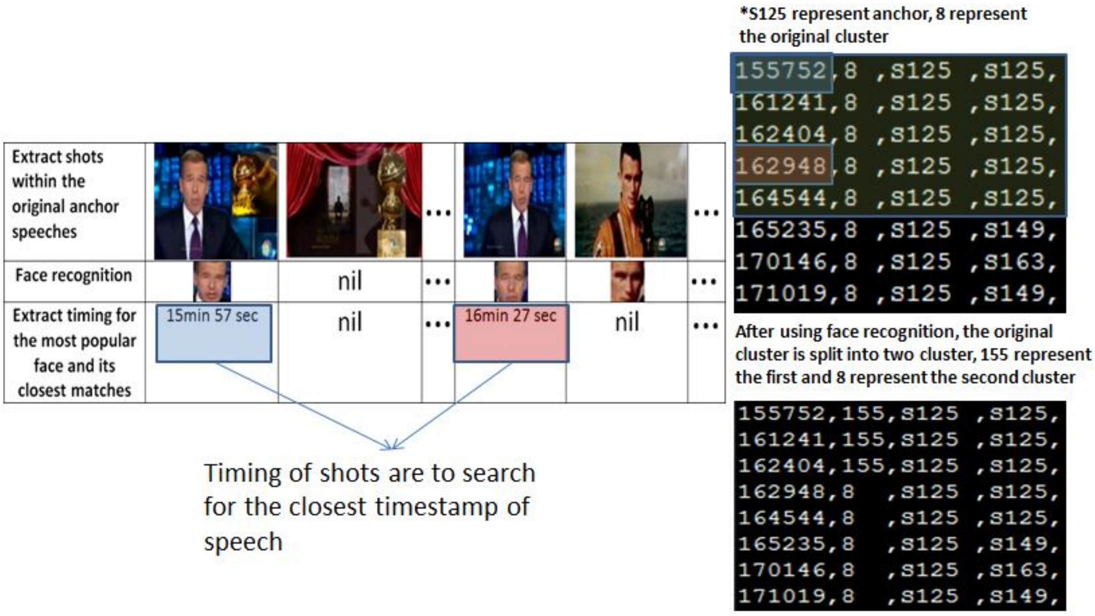


```
def faceCrop(imagePattern,boxScale=1):
    faceCascade = cv.Load('haarcascade_frontalface_alt.xml')

    imgList=glob.glob(imagePattern)
    if len(imgList)<=0:
        print 'No Images Found'
        return

    for img in imgList:
        pil_im=Image.open(img)
        cv_im=pil2cvGrey(pil_im)
        faces=DetectFace(cv_im,faceCascade)
        if faces:
            n=1
            for face in faces:
                croppedImage=imgCrop(pil_im, face[0],boxScale=boxScale)
                fname,ext=os.path.splitext(img)
                croppedImage.save(fname+'_crop'+ext)
                n+=1
                return (fname+'_crop'+ext)
```

Figure 4.15: pythonscript on face detection

- **cv.load()**: loads the training data haarcascade_frontalface_alt.xml for comparing.
- **Pil2cvGrey()**: convert the image to greyscale image.
- **DetectFace()**: produce the image with a bounding box around the faces if there is any.
- **imgCrop()**: produce the cropped image of the faces detected in the existing image.

4.4.2 Topic Boundary Detection using Face Recognition


Extract shots within the original anchor speeches
Face recognition
Extract timing for the most popular face and its closest matches
nil
nil
15min 57 sec
16min 27 sec
nil
Timing of shots are to search for the closest timestamp of speech

*S125 represent anchor, 8 represent the original cluster

```
155752,8 ,s125 ,s125,
161241,8 ,s125 ,s125,
162404,8 ,s125 ,s125,
162948,8 ,s125 ,s125,
164544,8 ,s125 ,s125,
165235,8 ,s125 ,s149,
170146,8 ,s125 ,s163,
171019,8 ,s125 ,s149,
```

After using face recognition, the original cluster is split into two cluster, 155 represent the first and 8 represent the second cluster

```
155752,155,s125 ,s125,
161241,155,s125 ,s125,
162404,155,s125 ,s125,
162948,8   ,s125 ,s125,
164544,8   ,s125 ,s125,
165235,8   ,s125 ,s149,
170146,8   ,s125 ,s163,
171019,8   ,s125 ,s149,
```

Figure 4.16: Process of detecting topic boundary using audio-visual approch

For each topic boundary, shots that are within the duration of anchor's speech lines are selected for face recognition. Shots with face extracted out were sent for comparison. For each face, a comparison with subsequent faces is done using colour histogram. The most popular face is assumed to be the anchor. For each timestamp of the anchor's image, it was matched to the closest timestamp of anchor's speech line. This step ensures that visual shots will not come into conflict with audio segmented boundaries. Subsequently, the selected speech line was converted to topic boundary. Those speech lines after it were converted to the same topic boundary until another topic boundary was met.

The use of face detection does not guarantee high accuracy. To improve the accuracy of face recognition, we considered 30% cropping of the right side of the keyframe.

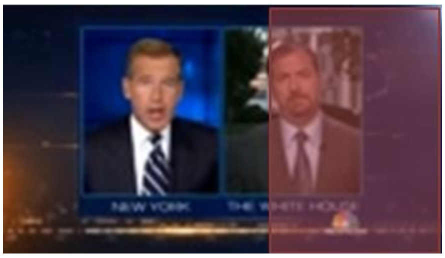
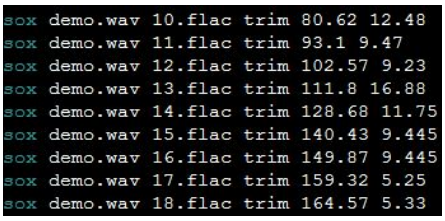


Figure 4.17: Improve accuracy by considering a smaller region of interest

4.5 Speech to Text

The segment file that consists of timestamps of speech lines are converted into audio files using SoundXchange(SoX) program. By providing the start time and the duration, SoX cut the original audio file into individual predefined audio segment. The format is also changed to $.flac$ as Google ASR could only handle $.flac$ format.

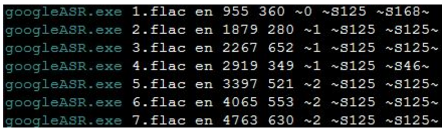


```
sox demo.wav 10.flac trim 80.62 12.48
sox demo.wav 11.flac trim 93.1 9.47
sox demo.wav 12.flac trim 102.57 9.23
sox demo.wav 13.flac trim 111.8 16.88
sox demo.wav 14.flac trim 128.68 11.75
sox demo.wav 15.flac trim 140.43 9.445
sox demo.wav 16.flac trim 149.87 9.445
sox demo.wav 17.flac trim 159.32 5.25
sox demo.wav 18.flac trim 164.57 5.33
```

Figure 4.18: SoX to cut audio file into individual audio segments

4.5.1 Google ASR

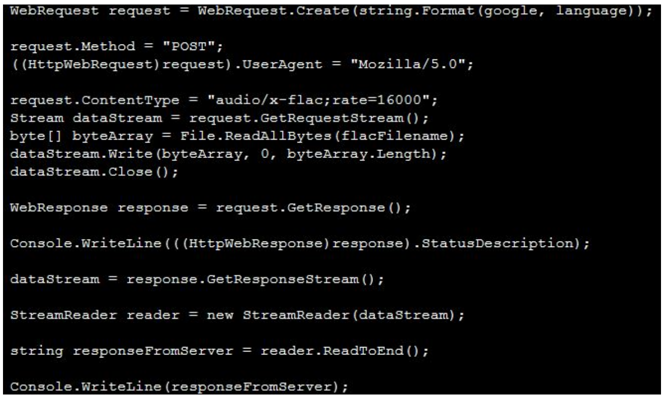
To execute large number of audio segments, *bat* file is used to call ASR executable program($C\#$ program). The language parameter *en* which represent English defines the spoken language of the audio. The other parameters are passed-over to be copied to its transcription.



```
googleASR.exe 1.flac en 955 360 ~0 ~S125 ~S168~
googleASR.exe 2.flac en 1879 280 ~1 ~S125 ~S125~
googleASR.exe 3.flac en 2267 652 ~1 ~S125 ~S125~
googleASR.exe 4.flac en 2919 349 ~1 ~S125 ~S46~
googleASR.exe 5.flac en 3397 521 ~2 ~S125 ~S125~
googleASR.exe 6.flac en 4065 553 ~2 ~S125 ~S125~
googleASR.exe 7.flac en 4763 630 ~2 ~S125 ~S125~
```

Figure 4.19: script to call googleASR executable program

To illustrate the communication to Google ASR, a detail *C#* code is shown below.



```
WebRequest request = WebRequest.Create(string.Format(google, language));

request.Method = "POST";
((HttpWebRequest)request).UserAgent = "Mozilla/5.0";

request.ContentType = "audio/x-flac;rate=16000";
Stream dataStream = request.GetRequestStream();
byte[] byteArray = File.ReadAllBytes(flacFilename);
dataStream.Write(byteArray, 0, byteArray.Length);
dataStream.Close();

WebResponse response = request.GetResponse();

Console.WriteLine(((HttpWebResponse)response).StatusDescription);

dataStream = response.GetResponseStream();

StreamReader reader = new StreamReader(dataStream);

string responseFromServer = reader.ReadToEnd();

Console.WriteLine(responseFromServer);
```

Figure 4.20: script to communicate with Google ASR

The request is sent to Google using POST method. Subsequently, the audio *.flac* file is written into bytes using ReadAllBytes() and sent to Google cloud service using dataStream.write(). The response returned with status that states the approval of the service. Lastly, the transcription is retrieved from Google cloud using getResponseStream().

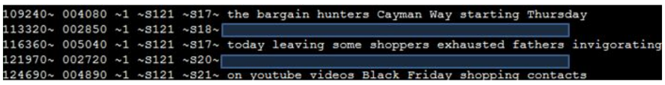
4.6 Text

After the transcription was completed, and topic boundary had been formalized. Text processing was done on each boundary to detect the topic. This whole process is accomplished using NLTK libraries.

4.6.1 Pre-processing

Remove empty line

Transcription from ASR is not perfect and often contains empty line as the ASR may not interpret certain speeches. The removal of these empty lines is essential as the transcription need to synchronize with the topic boundary properly.



```
109240~ 004080 ~1 ~S121 ~S17~ the bargain hunters Cayman Way starting Thursday
113320~ 002850 ~1 ~S121 ~S18~
116360~ 005040 ~1 ~S121 ~S17~ today leaving some shoppers exhausted fathers invigorating
121970~ 002720 ~1 ~S121 ~S20~
124690~ 004890 ~1 ~S121 ~S21~ on youtube videos Black Friday shopping contacts
```

Figure 4.21: Two emptylines within a boundary

Lemmatization/Stemming and Spelling Correction

Transcriptions from ASR often consist of spelling errors. To solve the problem, tokens are first extracted out by performing tokenization using NLTK function “nltk.word_tokenize(sentence)”. Subsequently, each token went through “SpellCheck()” function correcting grammatical errors. Lastly, transcriptions that contain different forms of a word such as “organize”, “organizes” and “organizing”; lemmatization and stemming were used to normalize these words to it common base form. This step is achieved using the wordnetLemmatizer().(*Wordnet is a large lexical database that contains English nouns, verbs, adjective, etc.*)

Remove meaningless words

Tokens that are two characters long or consist of special characters often bare little semantic meaning; they are removed to reduce inaccurate detection. The following regular expression is used to remove special characters:
re.sub(‘[ˆ$A - Za - z0 - 9$]+’,‘’,string)

4.6.2 Noun Words Extraction

After the pre-processing steps, POS tag is performed using “nltk.pos_tag(token)” which adds the speech tag for each token. Those tokens with noun tag are extracted out and stored in a list as they contain more semantic meaning than other speech tags. Subsequently, these noun words are stored in the database for news stories retrieval.

4.6.3 Topic Detection

In our proposed method, a dictionary consisting of ten categories (*politics and governance, military and war, entertainment, science and technology, culture and religion, history and geography, health, education, business*) are used. Each containing 2000 words is found by extracting keywords from Wikipedia.

For each noun word in the boundary, its frequency is calculated using the “freqDist()” and “fdist.keys()” function. Thereafter, it was compared with the topics in the dictionary. An alphabet index is included to speed up the searching process. The index jump to the specific alphabet based on the first letter of the query keyword. For instance, “obama” will cause the index to jump to the letter “O”. If we assumed the worst case scenario, the time complexity is O($\frac{1}{24}$) which is a small fraction of the total time. In order to assign more weights to words that

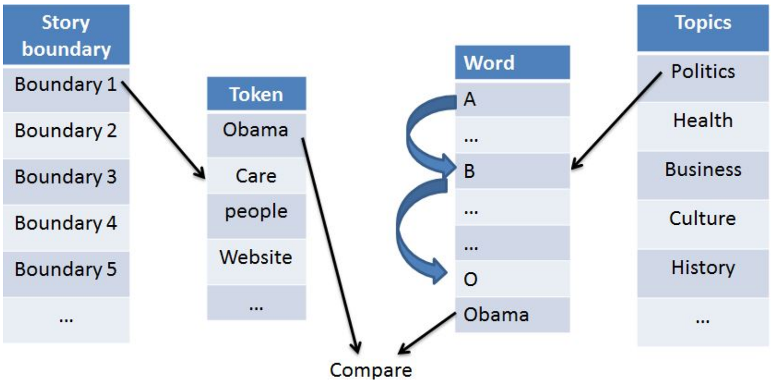


Figure 4.22: The matching algorithm

are more frequent in the topic boundary. A variation of TF-IDF is used. If we denote $tf_{t,q}$ as the term frequency of term t in topics boundary q, and d denote the topics in the dictionary. The $Freqscore$ is obtained as:

$$Freqscore = \sum_{t \in q \cap d} (1 + log tf_{t,q}) \tag{4.1}$$

However, the score is not enough to reflect the overall accuracy since each word that matched a topic carried the same weight regardless of how many topics it successfully matched. Therefore, for each word that appear in less number of topics, higher weight are given, than words that appear in more topics. If we denote N as the total number of topics in the dictionary, i as the subscript for each topic, $count$ in the form of 1 or 0 which indicate match or no match respectively. Then the $Topicscore$ is obtained as:

$$Topicscore = log(N - \sum_{i}^{N} count) + 1 \tag{4.2}$$

To reduce the impact of the weights, log_{10} is used for both $Freqscore$ and $Topicscore$. The score for both of them are added up and obtained as:

$$Score = Freqscore + Topicscore \tag{4.3}$$

In order to categorize the boundary into one of the topics in the dictionary, ranking of the score is used to determine it. The score in order is presented in Table 4.1.

Topic Boundary 1	
Topic in dictionary	Score
Topic 1	55%
Topic 2	30%
Topic 3	40%
Topic 4	22%

Table 4.1: Catergorizing boundary based on ranking

4.6.4 Story Headlines Detection

In each story boundary, POS chunking is carried out to extract the key phrases. In order to perform chunking, our proposed solution uses two methods mainly regular expression and a rule based classifier.

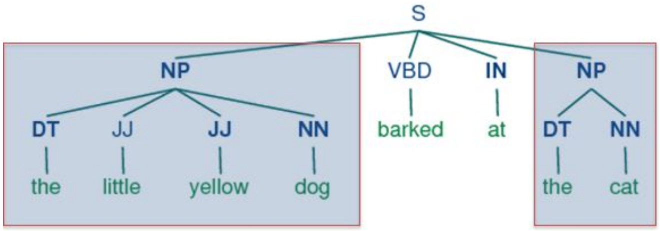
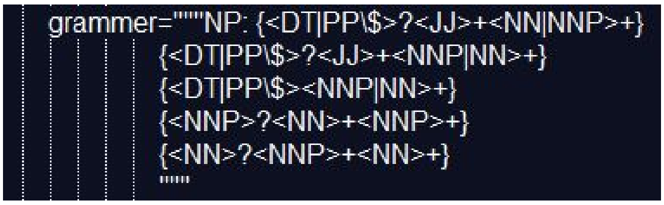


Figure 4.23: Finding Tag patterns

By using a regular expression E.g. $< DT >? < JJ.^* >^*< NN.^* > +$. This will chunked any sequence of tokens with zero or one determiner(DT), followed by zero or more adjectives (JJ), and lastly one or more nouns (NN).

Subsequently, the list of key phrases was compared with the top N noun tokens of its boundary, where N denotes a constant value. (*In this thesis, we consider N=3*) These tokens are derived by getting the highest frequency count using the freqDist() and fdist.keys() function. The key phrase that matched the most number of tokens was selected as headlines for the boundary. In the case where the number of matches is the same, the phrase with the longest length was chosen. An example of sample code to extract key phrases from sentences:

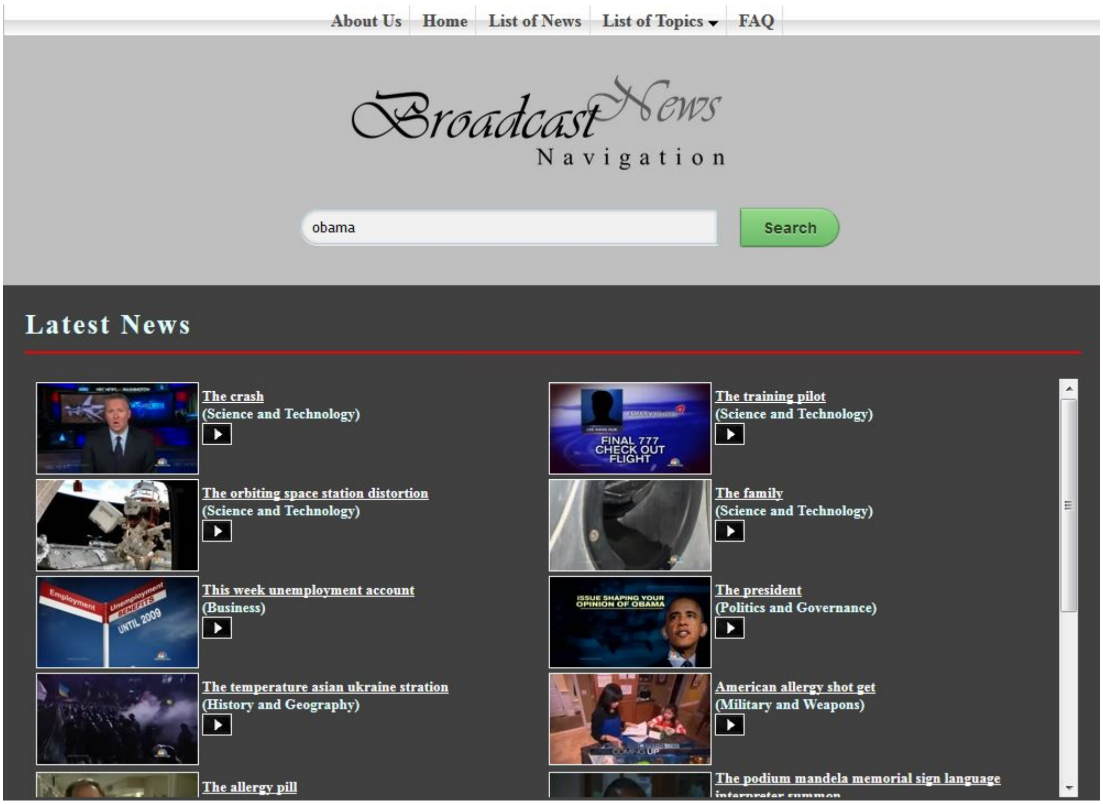


```
grammer="""NP: {<DT|PP\$>?<JJ>+<NN|NNP>+}
                {<DT|PP\$>?<JJ>+<NNP|NN>+}
                {<DT|PP\$><NNP|NN>+}
                {<NNP>?<NN>+<NNP>+}
                {<NN>?<NNP>+<NN>+}
                """
```

Figure 4.24: extract keyphrases using POS chunking

4.7 Broadcast News Navigation

To visualise the concept of the news segmentation, a web-based Broadcast News Navigation (BNN) is used. Jwplayer is embedded to a Webpage, and play when on click. Search and sort/filter functions are added to support retrieval of news stories. Other features such as text summarization, language translation, thumbnails scrolling are added to enhance user-friendliness.



The crash
(Science and Technology)

The training pilot
(Science and Technology)

The orbiting space station distortion
(Science and Technology)

The family
(Science and Technology)

This week unemployment account
(Business)

The president
(Politics and Governance)

The temperature asian ukraine stration
(History and Geography)

American allergy shot get
(Military and Weapons)

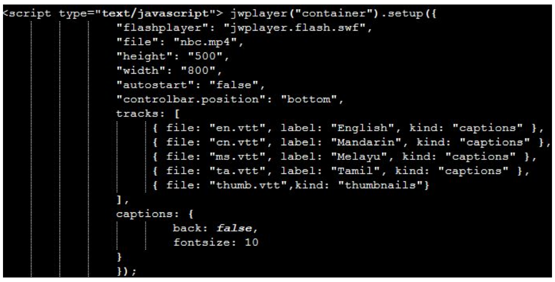
The allergy pill

The podium mandela memorial sign language interpreter summon

Figure 4.25: A web based Broadcast News Navigator

4.7.1 JWplayer

A following javascript to embed and customize the JWplayer in the HTML page:



```
<script type="text/javascript"> jwplayer("container").setup({
                "flashplayer": "jwplayer.flash.swf",
                "file": "nbc.mp4",
                "height": "500",
                "width": "800",
                "autostart": "false",
                "controlbar.position": "bottom",
                tracks: [
                     { file: "en.vtt", label: "English", kind: "captions" },
                     { file: "cn.vtt", label: "Mandarin", kind: "captions" },
                     { file: "ms.vtt", label: "Melayu", kind: "captions" },
                     { file: "ta.vtt", label: "Tamil", kind: "captions" },
                     { file: "thumb.vtt",kind: "thumbnails"}
                ],
                captions: {
                        back: false,
                        fontsize: 10
                }
                });
```

Figure 4.26: Script to embed JWplayer

- **height and width parameter**: to create sizes of the player in the website
- **autostart**: is turn to false to prevent the video from starting when the webpage is up
- **tracks**: consist of language translation file and thumbnails file.

In order for the video to skip to the specific timestamp issued by the user, jwplayer().seek() function is used. In JWplayer, durations are measured in seconds, thus jwplayer().seek(936) will skip to video position at 15mins and 36 sec.

Using the player

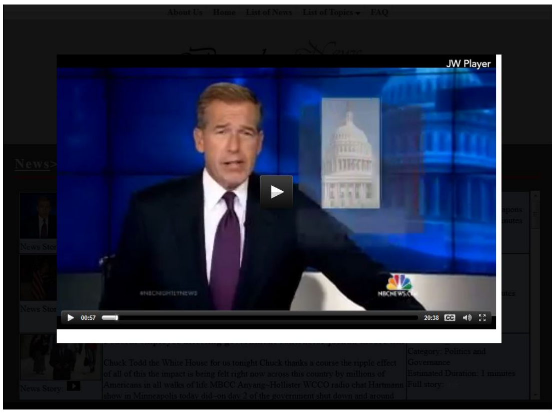
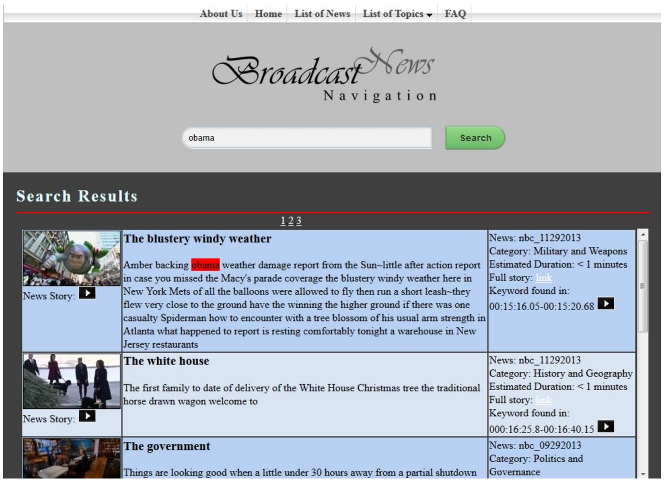


Figure 4.27: Playing the video using Jwplayer

When users click on the play icon ▶, JWplayer will be launched on a new pop out screen using "lightbox" effect. The use of pop out screen provides space saving as compared to a screen that stays in the mainpage in-regardless of whether the user uses it.

4.7.2 Search Engine

The search engine offers users a way to find content quickly. Users enter the query through the text box and the query was passed to the database. The database processes the query and returns the results. The results were sent to the web page for display using AJAX. As the number of news video increases, the number of records returned may increase which caused a tremendous amount of information presented to users. To rectify this issue, the results are split into pages of 5. Users can then select a particular page to visit.



News Story: ▶	**The blustery windy weather** Amber backing obama weather damage report from the Sun~little after action report in case you missed the Macy's parade coverage the blustery windy weather here in New York Mets of all the balloons were allowed to fly then run a short leash~they flew very close to the ground have the winning the higher ground if there was one casualty Spiderman how to encounter with a tree blossom of his usual arm strength in Atlanta what happened to report is resting comfortably tonight a warehouse in New Jersey restaurants	News: nbc_11292013 Category: Military and Weapons Estimated Duration: < 1 minutes Full story: link Keyword found in: 00:15:16.05-00:15:20.68 ▶
News Story: ▶	**The white house** The first family to date of delivery of the White House Christmas tree the traditional horse drawn wagon welcome to	News: nbc_11292013 Category: History and Geography Estimated Duration: < 1 minutes Full story: link Keyword found in: 000:16:25.8-00:16:40.15 ▶
	The government Things are looking good when a little under 30 hours away from a partial shutdown	News: nbc_09292013 Category: Politics and Governance

Figure 4.28: Information retrieval using the search engine

In the list of returned results, each containing a set of information; the information consists of:

- Duration for the news story
- The category it belongs
- The news video it belongs
- The headlines of the news story
- Summary of news story
- keyframe that represent the news story
- Timestamps of keyword appearing in the news story
- A play icon to play the video segment

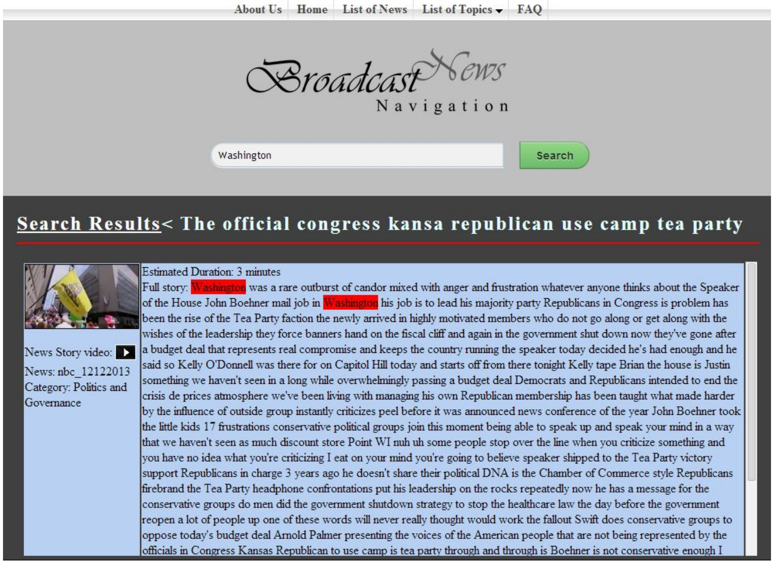
Timestamps of keyword

The results return a list of timestamps for each keyword that appears in the specific point of a news story. When the timestamp is clicked, the news video skip to the spot where the keyword was found.

Text summarization

A simple text summarization approach based on Luhn(1958) [20] was used in this project. It uses the frequency of each individual word in the topic boundary to calculate and extract the sentence that contains those frequent words. Using this approach, texts were summarized to a predefined number of sentences.
For each keyword found in the summary, it was highlighted in red. User can examine the highlighted text and determine if a particular news story is of his/her interests.

Full subtitle of a new story



Search Results< The official congress kansa republican use camp tea party

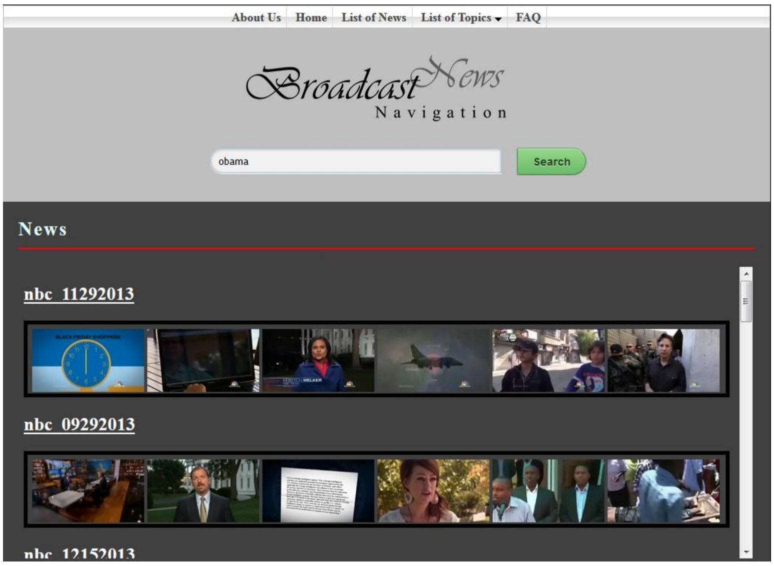
News Story video: ▶
News: nbc_12122013
Category: Politics and Governance

Estimated Duration: 3 minutes
Full story: Washington was a rare outburst of candor mixed with anger and frustration whatever anyone thinks about the Speaker of the House John Boehner mail job in Washington his job is to lead his majority party Republicans in Congress is problem has been the rise of the Tea Party faction the newly arrived in highly motivated members who do not go along or get along with the wishes of the leadership they force banners hand on the fiscal cliff and again in the government shut down now they've gone after a budget deal that represents real compromise and keeps the country running the speaker today decided he's had enough and he said so Kelly O'Donnell was there for on Capitol Hill today and starts off from there tonight Kelly tape Brian the house is Justin something we haven't seen in a long while overwhelmingly passing a budget deal Democrats and Republicans intended to end the crisis de prices atmosphere we've been living with managing his own Republican membership has been taught what made harder by the influence of outside group instantly criticizes peel before it was announced news conference of the year John Boehner took the little kids 17 frustrations conservative political groups join this moment being able to speak up and speak your mind in a way that we haven't seen as much discount store Point WI nuh uh some people stop over the line when you criticize something and you have no idea what you're criticizing I eat on your mind you're going to believe speaker shipped to the Tea Party victory support Republicans in charge 3 years ago he doesn't share their political DNA is the Chamber of Commerce style Republicans firebrand the Tea Party headphone confrontations put his leadership on the rocks repeatedly now he has a message for the conservative groups do men did the government shutdown strategy to stop the healthcare law the day before the government reopen a lot of people up one of these words will never really thought would work the fallout Swift does conservative groups to oppose today's budget deal Arnold Palmer presenting the voices of the American people that are not being represented by the officials in Congress Kansas Republican to use camp is tea party through and through is Boehner is not conservative enough I

Figure 4.29: Full subtitle of a particular news story

To provide a complete coverage of new stories, full subtitles for each new story are provided. User can also drill down to the keyword that is highlighted.

4.7.3 Sort and Filter options



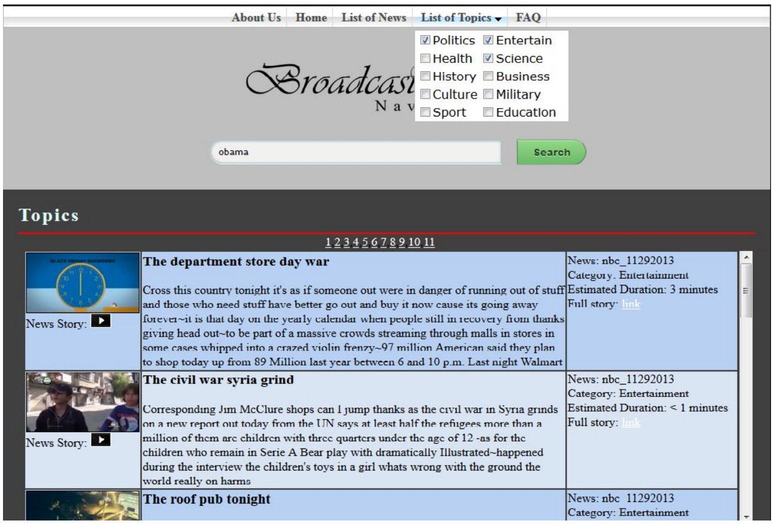
nbc_11292013

nbc_09292013

nbc_12152013

Figure 4.30: news are sorted according to each TV news programme

In order to enhance the user experience, users are able to view the news stories according to news programme or by topics. This allows them to view the records in the order they wish to view them. Besides, as the number of news increases, filtering options allow users to click on individual topics according to their preference. This allows them to only see the record that they wish to see.

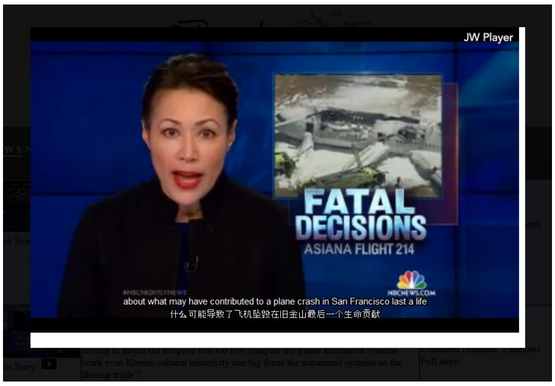


☑ Politics ☑ Entertain
☐ Health ☑ Science
☐ History ☐ Business
☐ Culture ☐ Military
☐ Sport ☐ Education

News Story: ▶	**The department store day war** Cross this country tonight it's as if someone out were in danger of running out of stuff and those who need stuff have better go out and buy it now cause its going away forever~it is that day on the yearly calendar when people still in recovery from thanks giving head out~to be part of a massive crowds streaming through malls in stores in some cases whipped into a crazed violin frenzy~97 million American said they plan to shop today up from 89 Million last year between 6 and 10 p.m. Last night Walmart	News: nbc_11292013 Category. Entertainment Estimated Duration: 3 minutes Full story: link
News Story: ▶	**The civil war syria grind** Corresponding Jim McClure shops can I jump thanks as the civil war in Syria grinds on a new report out today from the UN says at least half the refugees more than a million of them are children with three quarters under the age of 12 -as for the children who remain in Serie A Bear play with dramatically Illustrated~happened during the interview the children's toys in a girl whats wrong with the ground the world really on harms	News: nbc_11292013 Category: Entertainment Estimated Duration: < 1 minutes Full story: link
	The roof pub tonight	News: nbc 11292013 Category: Entertainment

Figure 4.31: news are filter according to each topic

4.7.4 Language Translation

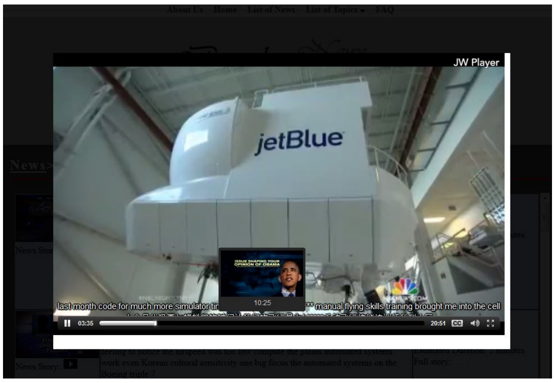


about what may have contributed to a plane crash in San Francisco last a life
什么可能导致了飞机坠毁在旧金山最后一个生命贡献

Figure 4.32: Translation embedded in the captions

In this thesis, language translation is done by passing the entire transcription to Google Translation. Four languages are chosen mainly English, Chinese, Melayu and Tamil respectively. The converted transcription is then written in VTT format to be embedded to the Jwplayer

4.7.5 Thumbnails Scrolling



last month code for much more simulator tir ** manual flying skills training brought me into the cell

Figure 4.33: Embedded thumbnails in Jwplayer

Thumbnails are small size images which are used to search for a particular scene in the news story. It was embedded in the JWplayer which form a series of images along the continuous video. These embedded thumbnails benefit users that want to jump the video to a specific point, by providing a preview of the frame. For each thumbnail that are processed out using FFmpeg, they are written in VTT format and embedded to JWplayer.

Chapter 5

Experiment and Results

5.1 Experiment Setup

The datasets had been specifically collected and annotated for this task. It consists of ten broadcast news, coming from NBC (National Broadcast Corp) news; an America broadcast News Corporation and CNA (Channel News Asia) news; a Singapore broadcast News Corporation. In each news program, there is only one anchor speaker. The length of each broadcast news program varies from 15 minutes to 30 minutes depending on the broadcast news network. Each news video consists on average about 300 shots and 10 scenes. In this experiment, the audio track is sampled at 16 kHz in one channel and converted to mono-channel audio before processing. For each video, the entire process took about 30 to 45 minutes to complete and uploaded to the database. To ensure an unbiased experiment, two volunteers are chosen to manually split the ten broadcast news video into news boundaries. For each boundary, they are to classify them into individual topics based on ten topics provided. In the case where both volunteers classified the boundary differently, a third volunteer is chosen to pick out one of the two topics selected by both of them. That topic will be used to classify the boundary.

5.1.1 Evaluation Measures

Topic boundaries evaluation

To evaluate topic segmentation, automatic comparison to manual segmented stories is done to extract the recall and precision scores.

$$\text{Recall} = \frac{\text{Correct detection}}{\text{Total number of boundaries detected}} \tag{5.1}$$

$$\text{Precision} = \frac{\text{Correct detection}}{\text{Correct detection} + \text{false detection}} \tag{5.2}$$

$$\text{F1-measure score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5.3}$$

To prove the concept of using multiple features provide a more reliable and accurate story boundaries. Evaluation will be done with one feature remove at a time, to estimate the loss of accuracy in terms of F1-measure score.

Topic Detection Evaluation

The reference text (ground truth) was used to compare with the automatic generated transcription using the word error rate (WER).In this experiment, word error rate(WER) is calculated as:

$$\text{WER} = \frac{\text{S+D+I}}{\text{N}} \tag{5.4}$$

$$\text{Accuracy} = \text{1- WER} \tag{5.5}$$

$$\text{Correct} = \frac{\text{N-D-S}}{\text{N}} \tag{5.6}$$

To evaluate the topic detection, the generated topics by ASR were compared with the reference topic to extract recall and precision scores.

$$\text{Recall} = \frac{\text{Number of correctly assigned topics}}{\text{Number of correct topics detects}} \tag{5.7}$$

$$\text{Precision} = \frac{\text{Number of correctly assigned topics}}{\text{Total number of topics}} \tag{5.8}$$

$$\text{F1-measure score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5.9}$$

The F1-measure score of each topic detection will be compared to the WER to analyse a relationship between them.

5.2 Results

Below are the summaries of results for ten news videos. Details for individual tests are included in the appendix.

5.2.1 Results for topic boundary

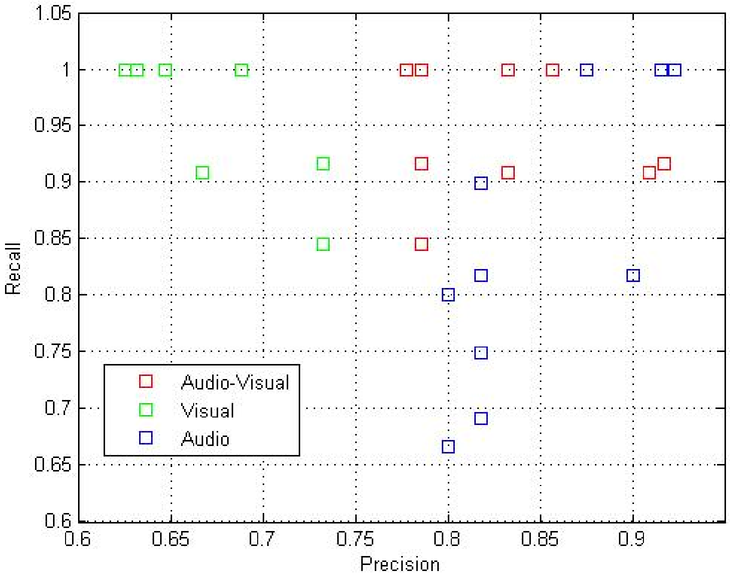


Figure 5.1: Comparison of each feature for 10 news video

Summary of recall, precision and F1 score:

	Audio-Visual	Visual only	Audio only
Recall	0.936	0.954	0.879
Precision	0.829	0.65	0.773
F1 score	0.879	0.85	0.483

Table 5.1: Summary of score for topic boundary evaluation of each category

	Goyal [1] (Visual)	Misra [31] (Text)	Hare [32] (Visual)	Proposed Method (Audio-Visual)
Recall	0.497	0.300	0.473	0.936
Precision	0.750	0.700	0.493	0.829
F1 score	0.600	0.420	0.483	0.879

Table 5.2: Comparison with the state-of-art methods

5.2.2 Results for topic detection

Summary of WER results:

	WER	Correct	Acc.
NBC news	0.340	0.690	0.659
CNA news	0.547	0.511	0.453

Table 5.3: WER results of both NBC and CNA news

Summary of recall, precision and F1-score:

	Recall	Precision	F1 score
NBC news	0.628	0.464	0.532
CNA news	0.531	0.441	0.474

Table 5.4: Score results of both NBC and CNA news

Comparison of WER against F1score for 10 records is shown in the figure below:

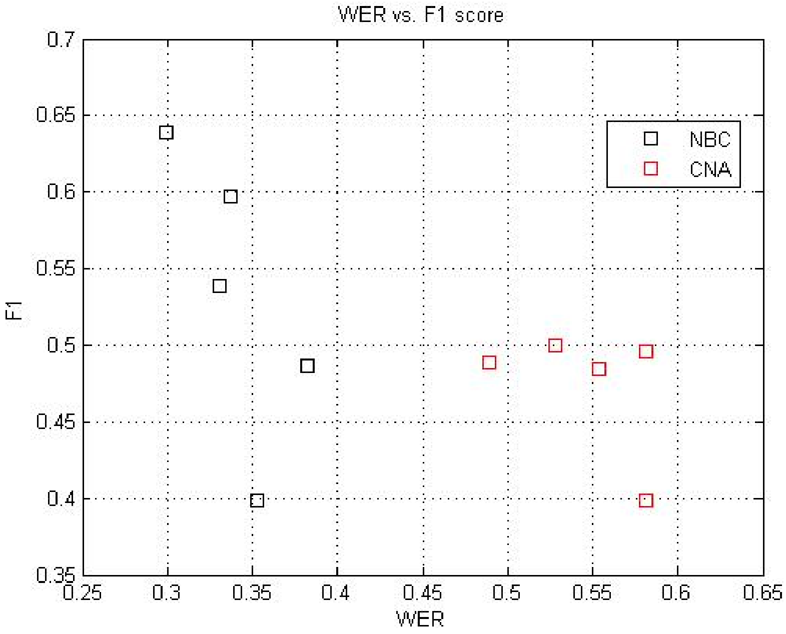


Figure 5.2: Comparsion between NBC and CNA news

5.3 Evaluation

5.3.1 Evaluation of Results (Topic boundary)

The above table provides the overall performance of topic segmentation algorithm. The results show that all the categories had score above 75% for its F1-score. However, this figure is not directly related to the usability of the system. We need to consider the Recall and the Precision in assessing how much impact each of its score will have effect on the results.

Recall considered the number of true positives that were found. In the three categories, audio method has the lowest recall score. This can be observed in figure 5.1 where audio points are mostly clustered at the bottom right. This is due to high false negatives. False negative relates to cases where a topic boundary is missed.boundary which resulted in users watching unrelated news before reaching their news. Audio method that has a low recall score clearly does not fulfil our objective to provide usability to our users.

Precision considered whether the numbers of return boundaries are true positive. In the three categories, visual method has the lowest precision score. This can be observed in figure 5.1 where visual points are mostly clustered at the top left. This is due to high false positives. False positive relates to cases where a video shot is incorrectly classified as boundary. As a result, the boundary is split into smaller boundaries. In terms of usability, it has no direct impact on users. However, smaller boundaries resulted in a smaller chunk of texts, which has direct impact on topic detection. One solution is to increase the threshold for colour similarity. However, that may result in more false negative boundaries. Therefore, visual alone clearly does not help in producing good topic boundaries.

In contrast, audio-visual method achieved promising results. In table 5.2, the results have shown that the proposed method improved greatly from the existing segmentation methods that applied only one feature. The use of the hybrid approach does have a significant influence over the score. For instance, the high recall score is due to the addition of the visual feature which overcomes limitations of the audio method. Likewise, the high precision score is due to the use of audio as the dominant feature which restrict the use of visual features to just within the anchor's speeches, therefore, reduces the error rate.

5.3.2 Evaluation of Results (Topic Detection)

In table 5.3, the results have shown that both NBC and CNA news performed averagely in WER with a score of 0.340 and 0.547 respectively. This is probably due to the limitation of Google ASR, where the transcription consists of numerous spelling errors and insertions/deletions of words. The WER for CNA videos is remarkably high (*higher score represents poorer accuracy*), with more than half of the texts in the transcription were different from the ground truth. This is due to Google ASR not able to recognise the Singapore accent clearly. This resulted in missed lines and high number of spelling errors.

Table 5.4 summarizes the results for both NBC and CNA news. In all scoring, both performed averagely. In comparison, NBC performed much better than CNA in identifying correct topics. These results seem to suggest a correlation between the WER and accuracy of topic detection. The higher the WER score, the lower the accuracy rate. The effect is shown in figure 5.2. The degradation in accuracy can be attributed to the number of spelling errors in the text. The topics are determined by comparing the words with a dictionary. Because the dictionary is unable to recognise spelling error, it is likely to detect wrong topic for the topic boundary based on only correct words.

However, we can observe for some news video with high WER, it has a higher F1-score than videos with lower WER. This phenomenal is due to the algorithm that only matches the existing correct words to the dictionary. It does not concern words that are missing, which is a component for scoring in WER. Therefore, a WER score with high number of *Insertion* does not necessary equate to a low accuracy score for topic detection.

To sum up, the high score of WER is mostly due to the limitation of current ASR technologies which were still not able to recognise the text clearly. Through the results for CNA news, we could also clearly observed that ASR only works well with its own native accent. On the other hand, the algorithm used for matching topics only concern words that are only available in the boundary. Therefore, topic detection may not accurately portray the relationship with WER well. Further improvement could have been done for better detection of topics.

Chapter 6

Conclusions and Future work

6.1 Summary of achievement

In this thesis, a fully automated broadcast news system was developed. From the stage of video downloading, extraction and analysis of each feature to media generation, no human intervention was required. A user simply chooses a particular news video to segment, and every part will be accomplished automatically by the proposed system.

The use of an agile approach had helped to improve the iteration processes in this thesis. In the first iteration, the focus was on using shot change to detect story boundaries, but each shot was far too small to have any visual meaning and it was difficult for indexing. In the second iteration, speaker diarization was used. The detection of story boundaries was fairly accurate, but improvement could be done to reduce the number of missed boundaries. In the third iteration, an audio-visual approach was implemented for story boundaries. In addition, the use of text for topic-detection was also implemented.

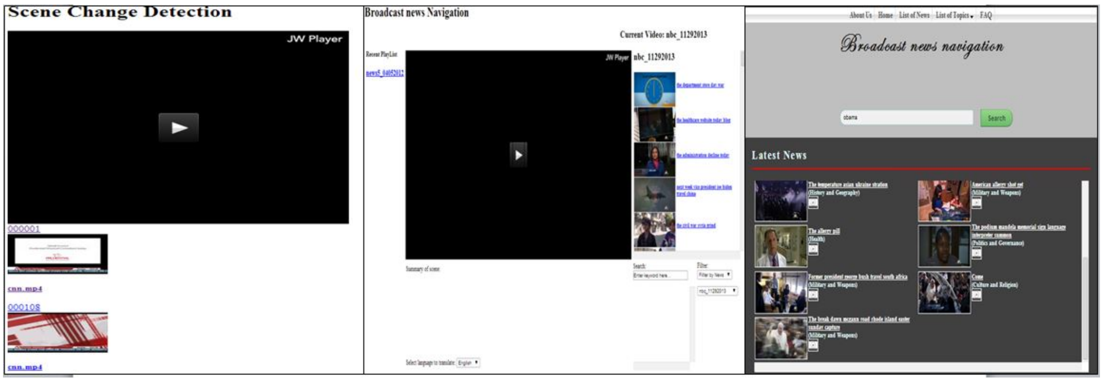


Figure 6.1: Change of web interface in each iteration

The results had proved that the combined audio-visual method achieved a high F1-score. This shows that the hybrid approach helped to overcome each of its own weakness and did not suffer the poor accuracy rate of other systems that relied only on one kind of feature.

In topic detection, POS tagging and chunking was successfully implemented to produce key phrases that represent the headlines of a news story. In addition, these news stories were categorized into different topics using a variation of IF-IDF for scoring. Third party libraries such as NTLK and wordnet provide natural language processing capabilities which were vital for our algorithm. The results had shown that topic detection achieved satisfactory accuracy where further improvements could have been done.

Our system followed a pipe and filter architecture, where each component was implemented in a modular fashion. This allows one to reuse any component from the system. In the future, re-usability of certain modules for other video segmentations is highly possible.

6.2 Future Directions

Due to the large scale of the system, many areas are not fully examined and there are some limitations in the current system. In this thesis, future research should explore following directions: 1) improvement on topic segmentation; 2) improvement on topic detection and 3) how to improve usability by data mining are briefly discussed in the following section.

6.2.1 Topic Segmentation

The audio-visual method used for topic segmentation, though perform decently, still suffered from missed and false detection. Besides, both audio and visual have the limitation of requiring a structured domain to perform. It is not able to resolve challenging segmentation problems such as lecture videos or *Tech Talk*. These videos often have low number of shot changes and speaker change detection is not usable since there is only one speaker. Unlike broadcast news, where the anchor speaks in a structured manner, these videos often contain spontaneous speeches by the speaker that are harder to segment. Topic segmentation by text could be used to resolve this issue. Text segmentation such as text tiling (*covered in chapter 2*) identifies the current topic of discussion and detects a topic change when texts are dissimilar in later boundaries. Generally, text segmentation is an unstructured way of dividing texts into different boundaries, and it is effective across many kind of segmentation problems.

6.2.2 Topic Detection

In this thesis, topic detection using a static dictionary is applied to identify topics for news boundaries. However, news of different kinds often consists of words that do not exist in the dictionary. Therefore, the current dictionary may not sustain in detecting topics with good accuracy. Currently, the solution was to add in news words extracted from Wikipedia manually. This increases the number of keywords, and a higher chance of matching the boundary. Another solution is to implement topic modelling. Topic modelling such as LDA (*discussed in chapter 2*) trained on news data and converts it into a dynamic dictionary. The larger amount of news data trained, the better the accuracy of the news topic. However, the use of LDA is highly dependent on the ASR. Words of the same meaning can appear in different forms, even after pre-processing, the dictionary may still store some of these words. Furthermore, large amount of news data are need for training in order to be effective. In terms of time and resources, we may not handle large amount information at the moment. Lastly, both the current method and topic modelling suffered from naming small boundaries. Small boundaries often consist of limited keywords, which often led to wrong topic detection. To resolve this problem, small boundaries could be removed, but at the expense of merging topics into one.

6.2.3 Mining of Multimedia Data

In this thesis, the main focus is mainly on segmentation and detection of topics. The topics generated for each boundary are common across all users. Although this approach improves searching, users are required to manually search for news that he/she interested. The use of data mining to predict the preference of users will greatly enhance user experience. In the simplest approach, we could alias with Google by retrieving user patterns, and present different topics to suit each user. Another approach is to rank the topics selected by users. The topic that was selected by the most number of users will be presented in the main page. In addition, the use of mining can help analysts to discover trends of topics that interest users, and therefore, could coordinate with news station to produce more of such news in the future.

Bibliography

[1] A.Goyal, P.Punitha, F.Hopfgartner, and J.M.Jose. Split and merge based story segmentation in news videos. In M.Boughanem, editor, *ECIR*, pages 766–770. Springer.Verlag, 2009.

[2] D.Morey A.Merlino and M.T. Maybury. Broadcast news navigation using story segmentation. In *ACM multimedia: 381-391*, 1997.

[3] F. Bellard. Ffmpeg. URL `http://www.ffmpeg.org/`.

[4] J.S. Boreczky and L.A. Rowe. Comparison of video shot boundary detection techniques. In *Storge and Retrieval for Image and Video Database(SPIE) 1996:170-179*, 1996.

[5] A. Brun and K. Sma. Experiment analysis in newspnews topic detection. In *SPIRE*, pages 55–64, 2000.

[6] M. Daneshi and M. Yu. Broadcast news story boundary detection using visual, audio and text features. In *ACM Multimedia*, 2013.

[7] D.Axmark and M.M.Widenius. *MySQL 5.5 Reference Manual*. MySQL Documentation Team, 5.1 edition, 1997.

[8] H. Dennis and A. Bechtel. Web site use and news topic and type. *JSMC Quarterly*, 79:73–86, 2002.

[9] S. Dharanpragada, M.Franz, J.S. McCarley, S. Roukos, and T.Ward. Story segmentation and topic detection in the broadcast news domain. In *in Proceedings of the DARPA Broadcast News Workshop (pp. 65-68)*, 1999.

[10] E.Avramidis and M.Melero. A richly annotated, multilingual parallel corpus for hybrid machine translation. In Nicoletta Calzolari and Khalid Choukri, editors, *LREC*, pages 2189–2193. European Language Resources Association (ELRA), 2012.

[11] E.Chen. Introduction to latent dirichlet allocation, Aug 2011. URL `http://blog.echen.me/2011/08/22/`.

[12] M.D. Fabro and L.Boszormenyl. State of the art and future challenges in video scene detection: a survey. In *Multimedia Syst. 19(5):427-454*, 2013.

[13] F.Morbini, K.Audhkhasi, K.Sagae, and R.Artstein. Which asr should i choose for my dialogue system? In *SLT*, pages 49–54, 2012.

[14] E. Alves G. Dias and C. Nunes. Topic segmentation: How much can we do by counting word and sequences of words. 2005.

[15] G.Gormley. Scene break detection & classification in digital video sequences. Technical report, Dublin City University, 1999.

[16] G.Salton and C.Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 5:513–523, 1988.

[17] M.A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 1:33–64, 1997.

[18] H.Fallon, A.Lattre, J.Bilien, A.Daoud, M.Gautier, and C.Stenac. VLC, version 2 edition, 2004.

[19] H.Jiang and T.Lin. Video segmentation with the assistance of audio content analysis. In *IEEE International Conference on Multimedia and Expo (III)*, 2000.

[20] H.P.Luhn. Automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165, 1958.

[21] O. Hyde. Big data solutions: Intelligent agents find meaning of text, Jan 2013. URL `http://www.ai-one.com/2013/01/18/`.

[22] K.Yamamoto I. Ide and H.Tanaka. An automatic video indexing method based on shot classification. In *Systems and Computers in Japan 32(9):32-41*, 2001.

[23] L.Y. Duan J. Wong and Q. Liu. A multimodal scheme for program segmentation and representation in broadcast video streams. *IEEE Transactions on Multimedia*, 3:393–408, 2008.

[24] K.Olmstead J.Sasseen and A.Mitchell. Digital: As mobile grows rapidly, the pressures on news intensify, 2013. URL http://stateofthemedia.org/2013/.

[25] J.Zetzsche. Let's talk: Trandos and the google translator toolkit. *The ATA Chronicle*, page 34, 2009.

[26] Z.A Khalaf. Automatic identification of broadcast news story boundaries using the unification method for popular nouns. In *Computer Science and Information System pp.577-584*, 2013.

[27] K.Markov and S.Nakamura. Never-ending learning system for on-line speaker diarization. In *ASRU 2007:699-704*, 2007.

[28] Lexalytics. Text analytics-phrase & theme extraction, Aug 2012. URL http://www.angross.com.

[29] *JW Player for Flash and HTML5*. Longtail video, 5.3 edition, October 2010.

[30] M.Huijbregts and D.A. Leeuwen. Diarization-based speaker retrieval for broadcast television archives. In *Interspeech 2011: 1037-1040*, 2011.

[31] H. Misra, F. Hopfgartner, and A. Goyal. Tv news story segmentation based on semantic coherence and content similarity. In *MMM2010:347-357*, 2010.

[32] N.Hare, A.F. Smeaton, C.Czirjek, N.Conner, and N.Murphy. A generic news story segmentation system and its evaluation. In *ICASSP*, pages 1028–1031, 2004.

[33] N.Seo. Opencv haartraining (rapid object detection), October 2008. URL http://note.sonots.com/SciSoftware/haartraining.html.

[34] Y. Park, S. Patwardhan, and S.C. Gates. An empirical analysis of word error rate and keyword error rate. In *Interspeech:2070-2073*, 2008.

[35] P.Wagner. Face recognition with python, July 2012. URL http://www.bytefish.de.

[36] G.V. Rossum. Python. URL https://www.python.org/.

[37] S.Bird, E.Klein, and E.Loper. *Natural Language Processing with Python.* OReilly Media, Inc, 2009.

[38] S.Meignier and T.Merlin. Lium spkdiarization: An open source toolkit for diarization. In *INTERSPEECH*, 2005.

[39] S.Panem, R.Bansal, and M.Gupta. Entity tracking in real-time using subtopic detecion on twitter. In *ECIR*, 2013.

[40] S.Shan. 15 awesome examples to manipulate audio files using sound exchange (sox), May 2009. URL `http://www.thegeekstuff.com/2009/05/`.

[41] V.Szalvay. *An Introduction to Agile Software Development*. Anube Technologies, http://www.danube.com, November 2004.

[42] M. Watson. Pos tagging, May 2012. URL `http://www.markwatson.com/opensorce/`.

[43] D.C Wong. *Speaker Diarization - "Who spoke when"*. PhD thesis, Queensland University of Technology, 2012.

[44] Y. Zhai and A.Yilmaz. Story segmentation in news videos using visual and text cues. In *CIVR*, 2005.

Appendices

.1 Evaluation for Boundaries detection (VisualAudio method

Topic boundary detection				
	True Positive	False Postive	False Negative	True Negative
Video 1: Human-Segment: 11 stories detected, 337 video shots				
AutoSegment	10	1	1	325
Video 2: Human-Segment: 13 stories detected, 267 video shots				
AutoSegment	11	3	2	251
Video 3: Human-Segment: 12 stories detected, 256 video shots				
AutoSegment	12	2	0	242
Video 4: Human-Segment: 10 stories detected, 288 video shots				
AutoSegment	9	2	1	276
Video 5: Human-Segment: 11 stories detected, 374 video shots				
AutoSegment	11	3	0	360
Video 6: Human-Segment: 10 stories detected, 307 video shots				
AutoSegment	10	2	0	295
Video 7: Human-Segment: 12 stories detected, 310 video shots				
AutoSegment	11	1	1	297
Video 8: Human-Segment: 7 stories detected, 265 video shots				
AutoSegment	7	2	0	256
Video 9: Human-Segment: 11 stories detected, 241 video shots				
AutoSegment	10	2	1	228
Video 10: Human-Segment: 12 stories detected, 350 video shots				
AutoSegment	11	3	1	335

Reference	Boundary	Not Boundary
Boundary	102	7
Not Boundary	21	

Table 1: Summary of score for audio-visual method

.2 Evaluation for Boundaries detection (Visual)

Topic boundary detection				
	True Positive	False Postive	False Negative	True Negative
Video 1: Human-Segment: 11 stories detected, 337 video shots				
AutoSegment	11	5	0	321
Video 2: Human-Segment: 13 stories detected, 267 video shots				
AutoSegment	11	4	2	250
Video 3: Human-Segment: 12 stories detected, 256 video shots				
AutoSegment	11	4	1	239
Video 4: Human-Segment: 10 stories detected, 288 video shots				
AutoSegment	10	7	0	272
Video 5: Human-Segment: 11 stories detected, 374 video shots				
AutoSegment	11	6	0	357
Video 6: Human-Segment: 10 stories detected, 307 video shots				
AutoSegment	10	6	0	291
Video 7: Human-Segment: 12 stories detected, 310 video shots				
AutoSegment	11	4	1	294
Video 8: Human-Segment: 7 stories detected, 265 video shots				
AutoSegment	7	8	0	250
Video 9: Human-Segment: 11 stories detected, 241 video shots				
AutoSegment	10	5	1	225
Video 10: Human-Segment: 12 stories detected, 350 video shots				
AutoSegment	12	7	0	332

Reference	Boundary	Not Boundary
Boundary	104	5
Not Boundary	56	

Table 2: Summary of score for visual method

.3 Evaluation for Boundaries detection (Audio)

Topic boundary detection				
	True Positive	False Postive	False Negative	True Negative
Video 1: Human-Segment: 11 stories detected, 337 video shots				
AutoSegment	9	1	2	325
Video 2: Human-Segment: 13 stories detected, 267 video shots				
AutoSegment	9	2	4	252
Video 3: Human-Segment: 12 stories detected, 256 video shots				
AutoSegment	9	2	3	242
Video 4: Human-Segment: 10 stories detected, 288 video shots				
AutoSegment	9	2	1	276
Video 5: Human-Segment: 11 stories detected, 374 video shots				
AutoSegment	9	2	2	361
Video 6: Human-Segment: 10 stories detected, 307 video shots				
AutoSegment	8	2	2	295
Video 7: Human-Segment: 12 stories detected, 310 video shots				
AutoSegment	12	1	0	309
Video 8: Human-Segment: 7 stories detected, 265 video shots				
AutoSegment	7	1	0	257
Video 9: Human-Segment: 11 stories detected, 241 video shots				
AutoSegment	11	1	0	229
Video 10: Human-Segment: 12 stories detected, 350 video shots				
AutoSegment	8	2	4	336

Reference	Boundary	Not Boundary
Boundary	91	18
Not Boundary	16	

Table 3: Summary of score for audio method

.4 Word error rate(NBC news)

	H	D	S	I	N	WER	correct	Acc
video 1	2289	513	498	80	3300	0.331	0.693	0.669
video 2	2262	484	465	132	3211	0.337	0.704	0.663
video 3	2256	644	549	126	3449	0.382	0.654	0.618
video 4	2315	594	513	101	3422	0.353	0.676	0.647
video 5	2119	397	410	69	2926	0.299	0.724	0.701

.5 Word error rate (CNA news)

	H	D	S	I	N	WER	correct	Acc
video 1	991	371	449	137	1811	0.528	0.547	0.472
video 2	1245	863	572	123	2754	0.581	0.465	0.419
video 3	1468	740	431	121	2639	0.489	0.556	0.511
video 4	1338	940	512	168	2790	0.581	0.479	0.419
video 5	1401	841	508	177	2750	0.554	0.509	0.446

.6 Topic detection (NBC news)

Main Topic						
	Reference	TP	FP	Recall	Precision	F1 score
Video 1	11	7	4	0.636	0.467	0.539
video 2	13	9	4	0.692	0.529	0.597
video 3	12	7	5	0.583	0.418	0.487
video 4	10	5	5	0.500	0.333	0.399
video 5	11	8	3	0.727	0.571	0.639

.7 Topic Detection (CNA news)

Main Topic						
	Reference	TP	FP	Recall	Precision	F1 score
Video 1	10	6	4	0.600	0.429	0.500
video 2	12	6	6	0.500	0.333	0.399
video 3	7	3	4	0.428	0.571	0.489
video 4	11	6	5	0.545	0.455	0.496
video 5	12	7	5	0.583	0.416	0.639