Utilization of Multimodal Interaction Signals for Automatic Summarisation of Academic Presentations

Keith Curtis

B.Sc.

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisors:
Prof. Gareth Jones
Prof. Nick Campbell (Trinity College Dublin)

June 2018

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

(Candidate) ID No.: 59847065

Date:

Contents

List of Figures

List of Tables

Utilization of Multimodal Interaction Signals for Automatic Summarisation of Academic Presentations

Keith Curtis

Abstract

Multimedia archives are expanding rapidly. For these, there exists a shortage of retrieval and summarisation techniques for accessing and browsing content where the main information exists in the audio stream. This thesis describes an investigation into the development of novel feature extraction and summarisation techniques for audio-visual recordings of academic presentations.

We report on the development of a multimodal dataset of academic presentations. This dataset is labelled by human annotators to the concepts of presentation ratings, audience engagement levels, speaker emphasis, and audience comprehension. We investigate the automatic classification of speaker ratings and audience engagement by extracting audio-visual features from video of the presenter and audience and training classifiers to predict speaker ratings and engagement levels. Following this, we investigate automatic identification of areas of emphasised speech. By analysing all human annotated areas of emphasised speech, minimum speech pitch and gesticulation are identified as indicating emphasised speech when occurring together.

Investigations are conducted into the speaker's potential to be comprehended by the audience. Following crowdsourced annotation of comprehension levels during academic presentations, a set of audio-visual features considered most likely to affect comprehension levels are extracted. Classifiers are trained on these features and comprehension levels could be predicted over a 7-class scale to an accuracy of 49%, and over a binary distribution to an accuracy of 85%.

Presentation summaries are built by segmenting speech transcripts into phrases, and using keywords extracted from the transcripts in conjunction with extracted paralinguistic features. Highest ranking segments are then extracted to build presentation summaries. Summaries are evaluated by performing eye-tracking experiments as participants watch presentation videos. Participants were found to be consistently more engaged for presentation summaries than for full presentations. Summaries were also found to contain a higher concentration of new information than full presentations.

Acknowledgments

I would first like to express my appreciation to my two supervisors, Professor Gareth Jones (of Dublin City University) and Professor Nick Campbell (of Trinity College Dublin). I have been lucky to have two such great supervisors.

I would like to thank Nick for his particular encouragement in the early days of my PhD studies as he encouraged me to think as a researcher, and for introducing me to the vast array of multimodal research in his lab in TCD.

I would like to thank Gareth for his continued support and advice and constant feedback. Gareth is always very supportive of his students and his advice has been invaluable, not just for my PhD thesis but for my career direction as a whole. Gareth's advice and feedback was essential, not just in the preparation of this thesis, but in the preparation of all my publications throughout the course of my research.

I would also like to thank my examiners Dr Kristiina Jokinen and Professor Noel O'Connor, and my independent chairperson Dr Alistair Sutherland for making my PhD defence a memorable experience. I also want to express to gratitude to my ADAPT and DCU colleagues, whose support and professionalism was essential in making my PhD studies such a memorable experience. I would also like to express my gratitude to all friends within our lab who assisted with frequent annotation requests, without which much of this work would not have been possible.

I would like to express my severe gratitude to my family who have supported and encouraged me all the way. I am eternally grateful to my parents, Anna and Brian, my sister Sarah and my niece Chelsea for all the years of encouragement and support.

Last but not the least, I am thankful to Science Foundation of Ireland for the research grant (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) to support this research.

Chapter 1

Introduction

Large archives of multimedia content are currently being created increasingly rapidly, with enormous levels of consumption of the resulting archives. Every minute in 2017, 4.1 million videos were viewed on *YouTube*, 70,017 hours of video were watched on *NETFLIX*, 40,000 hours of music were listened to on *Spotify* and 46,200 new posts were made to *Instagram*.[1] In this environment, it is a growing challenge to locate and browse content of interest in such multimedia archives - either in response to user queries or in freer or more informal exploration of content. Interactive browsing of multimedia archives to find relevant information can be extremely time consuming. While this is a challenging problem for multimedia including a video channel, this is particularly challenging for spoken content where a user must listen to the spoken audio track in a linear mode. This represents a growing challenge for accessing content considered relevant or of interest to users of multimedia archives where the significant information is in the audio of the multimedia file such as lectures, presentations etc., of which archives are growing rapidly.

Current methods for accessing and browsing digital video archives include scene type classification (Bosch et al., 2008) for named places, matching to low level features such as colours (Zhao and Grosky, 2002), or shape and textures (Rui et al., 1998; Chang and Smith, 1995). For finding relevant content, research has primar-

[1] `http://thebln.com/2017/03/happens-every-minute-internet-2017/`

ily focused on matching text queries against written metadata or transcribed audio (Chechik et al., 2008). These methods are limited by available low-level metadata descriptions and transcripts when browsing content such as lectures or presentations, where typically little information is available in the visual stream and most of the information exists in the audio track. The work described in this thesis contributes to addressing these limitations by classifying high-level paralinguistic features in such content, which can then be used for summarising lectures or presentations.

In order to address the problems of efficient engagement with this content, in this work we seek to classify paralinguistic features from within audio-visual presentations, and investigate the potential for the use of these high-level features for automatic summarisation of presentations.

In this chapter we introduce the subject of the thesis, along with the motivations of the work and main research objectives. An overview of the research problem is first introduced. Following this, the main objectives of this research are introduced. We then introduce the Research Questions which are the focus of this thesis. An overview of the structure of the remainder of the thesis then follows.

1.1 Motivation

There are a growing number of online archives of audio-visual presentations and lectures. Typically a user will watch these from the start in a linear fashion through to the end of the video. This serves the needs of a typical user, who may be learning the topic for the first time, or those of a user who may already know the start and end times in advance for the part(s) of the video they are interested in viewing. However, in a scenario where the user is already familiar with much of the discussion and only wishes to review a certain part of the video, or only wants to view the parts of the video considered most engaging and interesting, or wishes to browse to the key points as determined by the speaker as well as the content, or only wishes to get a general overview of the topic of discussion, current linear browsing methods are time consuming and inefficient.

Many users do not want to take the time to view full presentations and lectures, particularly if they are unsure as to whether the material is going to be of interest to them, or if much of the material is not of direct interest to them. Author-generated abstracts can be helpful in aiding the user to decide if the material is to be of benefit for them, however, most presentations and lectures are not accompanied by such abstracts. Even in cases where these are provided, the lecture or presentation may be too long for the user to commit to watching the presentation in its entirety. An alternative to viewing a video in its entirety is to engage with a summary of the key elements of the video.

Current video summarisation methods typically focus on the visual stream for event and salience detection. These summarisation methods do not cater well for material such as presentation and lecture videos where most of the information exists in the audio stream. In this regard, we consider the possibility of automatically classifying parts of presentations to high-level concepts familiar to humans, such as audience engagement and potential comprehensibility and emphasised speech.

We hypothesise that access to audio-visual lectures and presentations can be improved by automatically identifying areas considered to be the most engaging, emphasised or of most potential comprehensibility for the viewer. Automatically classifying areas of presentations considered most engaging and comprehensible for the audience could be used to implement techniques for the automatic generation of user summaries. These can give users a full overview of the presentation in question, allowing them to decide if that presentation is of interest to them. Based on that information, users can decide whether to watch the full presentation or to generate and watch a short summary of the presentation.

The automatic generation of presentation and lecture summaries would enable people to watch these to gain a quick overview of the material in question and see if it is going to be of interest and relevant to them. Thus less time would be wasted watching presentations and lectures not of interest or relevant to the user, while for relevant and interesting material the user could decide whether to watch the full presentation or to instead watch a longer, more in depth automatically generated summary of the material.

The personal motivation for the choice of features chosen for this research comes from the recording of the presentations. Having set out with the intention to develop a set of methods to improve summarisation of presentations, we recorded the data set used in this thesis as described in Chapter 3. During one of the presentations of this data set in particular there was obvious great interest / engagement from the audience. This led me to wonder if it would be possible to automatically detect such presentations which are greatly engaging for the whole audience and whether the speaking techniques of the presenter encouraged the interest. Following the work on engagement, the concept of comprehension, how much the audience can actually understand the content, and spoken emphasis, areas of speech specifically emphasised by the presenter of having specific importance, seemed the natural follow-up to the question of engagement / interest.

1.1.1 Example Use-Case 1

Emily is a researcher and wishes to watch recordings of recent conference presentations which she has missed. However, Emily is very busy and does not want to waste time watching presentations which are of no interest to her. There are a number of presentations which may potentially be of use to her work.

Using a new application based on the work described in this thesis, she are now able to automatically generate summaries of the presentations. Emily decides to generate short summaries of all the missed presentations she feels could potentially be of interest to her. Having watched the short presentation summaries, Emily then decides that one presentation in particular is of potential importance to her current work and she wishes to watch this presentation in full, but also decides that she would like a more in depth summary of two of the presentations without having the commit to watching the full talk, as they are of general interest but not crucial for her work.

Emily decides to generate two longer, in-depth summaries of the talks for which she does not want to commit to watching the full presentations. She then watches these longer summaries in addition to the full talk which she had committed to watching.

1.2 Overview

The main aim of the work described in this thesis is to investigate the creation of efficient and effective mechanisms to summarise audio-visual recordings where the significant information is in the audio stream. Current research on multimedia information retrieval has focused primarily on matching text queries against written metadata or transcribed audio (Chechik et al., 2008), on matching visual queries to low-level features such as colours (Zhao and Grosky, 2002; Niblack et al., 1993), textures and shape (Rui et al., 1998; Chang and Smith, 1995), and object recognition and person recognition (Marszalek and Schmid, 2007; Aslandogan and Yu, 1999; Cheng et al., 1998), in addition to scene type classification (Bosch et al., 2008; Szummer and Picard, 1998) - urban, countryside, and named individuals or places etc. (Huiskes et al., 2010; Xiao et al., 2010; Gallagher and Chen, 2009).

In the case of multimedia content where the information is primarily visual, the content can be represented by multiple key frames extracted from the video, using methods such as object recognition and facial detection. These are then matched against visual queries, and the retrieved videos are shown to the user using keyframe surrogates, with the user playing back those full videos that they believe to be relevant to their information need. Matching of the visual component of these queries is generally complemented by textual search against a transcript of any available spoken audio and any other metadata provided with the video (Lew et al., 2006).

The above methods are intended for content with a significant visual dimension, for which spoken content provides additional or complementary information to the visual information stream. However, for significant amounts of multimedia content, the information that they contain is not primarily visual, e.g. public presentations such as lectures, which largely focus on a single speaker talking at length on a single topic, or meetings, where multiple speakers discuss a range of previously selected issues. Text-based searching of a transcript of recorded speech may find a relevant recording, but there is currently a lack of support to assist users in finding content most likely to be engaging and interesting for them. To help users avoid large amounts of inefficient browsing of retrieved content, how might we assist them to decide which parts to review? In the context of multi-topic items, even identifying the most relevant recordings may be difficult due to the general challenge of searching content covering multiple topics. How might we help the user to locate the most significant items? In this research, we aim to extend the state of the art in multimedia search, taking the research in a new direction by utilising non-verbal communication signals in order to enable users to more efficiently access content relevant to their information needs contained in multimedia archives. We propose to fuse and analyse non-verbal audio and visual signals to provide features identifying information which will be of interest to searches looking for relevant content. Rather than standard work in Multimedia Information Retrieval (MIR), which investigates search methods to find relevant documents, our work focuses on investigation of individual relevant multimedia documents to explore methods to identify areas that are likely to be of particular interest to searchers.

These non-verbal communication signals include acoustic features such as *pitch*, *intensity*, *speech rate*, and *articulation rate*, and visual features such as *facial detection*, *object recognition*, *edge detection*, *optical flow*, and *colour histogram*. For this research, we investigate whether these non-verbal signals can be combined and fused to learn communicative social signals which could possibly serve as indicators of the social and communicative aspects of the interaction.

The research described in this thesis in based on conference presentations rather than meetings, with typically just one or two presenters speaking in front of a large audience. For this reason we conduct the research described here on a data-set of academic presentations, consisting of video of the presenter and corresponding video of the audience.

For classification of the features of engagement and comprehension we include behaviours of the audience during presentations such as movements, attendance at talks and the numbers facing forward towards the speaker or the presentation slides. We have not included post-presentation questions from the audience in this analysis, which are also indicative of engagement, primarily because the purpose of this work is to summarise the presentation itself by the most engaging, comprehensible parts. While questions from the audience following a presentation are useful indicators of engagement in itself, they often do not indicate the parts of a presentation in which audience participants were most engaged. Questions also do not form a part of summaries and thus have been excluded from the analysis.

1.3 Objectives

The objectives of this research project is to provide an advanced video summarisation method, which can be used to enhance the experience of users wishing to review academic presentations. As will be expanded upon in the following chapter, current automatic video summarisation techniques remain quite limited, particularly for viewing academic presentations and lectures, as most of the information in these aspects exists in the audio stream with very little in the way of visual stimuli, which are typically used for summarisation of video.

To accomplish this we investigate whether it is possible to classify higher level concepts than those typically classified for the summarisation of video, namely we investigate whether it is possible to build classifiers to identify the normally subjective terms of ‘good’ speaking techniques, audience engagement, intentional or unintentional speaker emphasis, and areas of potential comprehensibility for the audience. To this end we develop a dataset of academic presentations, recorded at an international conference related to speech.

We first aim to develop a technique to accurately classify areas of speech which humans regularly consider to be ‘good’ speaking techniques. While there exists a number of guidelines for public speakers on this topic, this remains a very subjective concept, with people regularly having their own ideas on what represents a ‘good’ speaker. As this is such a subjective topic, we engage human annotators to provide ratings for the presenter throughout the data set. We then seek to use these information annotations to predict audience engagement levels for such presentations. Human annotators label engagement levels throughout the video. By using automatically extracted low-level audio-visual features, we attempt to train a classifier to automatically predict speaker ratings and audience engagement.

We investigate the development of methods to identify areas of intentional or unintentional emphasis by the presenter, and to assess whether these areas show correlation with areas of high audience engagement, as identified by human annotators.

The logical next step in this line of research leads us to investigate potential audience comprehension of academic presentations. As engagement asks the question as to how much the audience are paying attention and mentally involved in the presentation, comprehension asks the question of how much the audience can actually follow and understand the material in question. While much of this is dependant on the prior experience and knowledge of the individual audience member, we hypothesise that a clear presentation structure together with clear, fluent speech on the part of the speaker, can aid audience members in their understanding of the material. We crowdsource human annotations of the dataset on this concept. Once again, by extracting multiple audio-visual features from each available modality, we attempt to train a classifier to classify areas of potential comprehensibility.

Following this initial experimental work, we aim to develop a novel video summarisation algorithm to summarise academic presentations by the content that a user is more likely to find engaging and comprehensible. This application directly addresses the motivation for this work, introduced in Section 1.1.

We provide a user-focused evaluation of our automatically generated video summaries. As the core part of this evaluation, we will request human participants to take part in eye-tracking evaluations of automatically generated presentation summaries and full presentations. With good, engaging, and comprehensible summaries, we would expect participants to keep their focus for longer periods than for full presentations, which we can evaluate from participants' focus. We would also expect that good summaries would result in participants being highly engaged as they watch the summaries. By comparing the number and duration of fixations we can evaluate whether participants are highly engaged for presentations summaries, and particularly whether or not they are more highly engaged than for full presentations. We provide further evaluation of summaries by crowd-sourcing questionnaires on presentation summaries and by polling participants on their ease of use and effectiveness for gaining a quick overview of the content of the video.

1.4 Research Questions

The preceding overview, which outlines the motives and primary objectives for this research leads to the following hypothesis: *Social signals on the part of a presenter / speaker can be utilised to train classifiers which can be used to identify concepts such as audience engagement, strengths of the presenter, intentional or unintentional emphasis of speech, and the speaker's potential to be comprehended / levels of understanding among the audience to academic presentations and lectures, and that such classifiers can be utilised to provide novel and effective features for summarisation of such video content.*

To explore the detailed elements of this hypothesis, we have identified the following five research questions to be addressed in this research:

In our first two research questions we investigate whether it is possible to provide a method for prediction of audience engagement and interest by analysing the speaking skills of the speaker in multimodal content. Previous work (Bednarik et al., 2012; Oertel and Salvi, 2013; Bonin et al., 2012; Gatica-Perez et al., 2005) in engagement detection has focused mainly on the annotation and detection of engagement in conversations. Other authors, (Whitehill et al., 2014; Jang et al., 2014) report work on engagement detection in multimodal learning environments. Other studies (Bohus and Horvitz, 2014; Corrigan et al., 2014) focused on engagement detection in human-robot conversation. In our work we study audience engagement with presentations given in academic conferences.

Meanwhile, previous studies on good public speaking techniques (Strangert and Gustafson, 2008; Strangert, 2007; Chen et al., 2014) focused on extracting the most influential, mainly acoustic, features associated with good public speaking skills. Other work (Liscombe et al., 2003) was performed on the detection of emotional speech. We study the features of speech that humans consider to be good speaking techniques and explore the training of a classifier for these features.

- Research Question 1 (RQ-1): Can we build a classifier to automatically rate the qualities of a good public speaker?
- Research Question 2 (RQ-2): Can we build a classifier to automatically predict the levels of audience engagement by utilising speaker-based and visual audience-based modalities?

Previous work has been performed on the automatic detection of emphasised speech in a unimodal context. (Chen and Withgott, 1992) studied the use of emphasis for automatic summarisation of a spoken discourse. Emphasised speech from one speaker was detected and summarisation excerpts were extracted with no noticeable differences with human extracted summarization excerpts. A Hidden Markov Model (HHM) was used to train the emphasis detector. (Arons, 1994) performed similar work, this time classifying the top 1% of speech pitch values as emphasised speech. Again, these speech extracts were found to provide a good basis for summarisation. Following this work, (Kennedy and Ellis, 2003) studied emphasis detection for characterisation of meeting recordings. In this work, five human annotators labelled 22 minutes of audio from the ICSI meeting corpus. To do so, transcripts were given to annotators to account for context, while annotators marked 'neutral' or 'emphasised' for each utterance. Pitch versus Time was extracted for each of the speakers in the meeting, and the corresponding mean and standard deviation were calculated. In cases where four or more human annotators agreed on emphasised speech, accuracy rates of 92% were achieved. In addition, the utterances found to be the most emphasised were rated by annotators as a good summarisation of the meeting recording.

In this thesis, we investigate correlations between fine-grained audience engagement and areas of emphasised speech. This will provide insights into the effects on the audience of emphasised speech during presentations. Positive correlations in this regard along with positive detections of emphasised speech may provide us with evidence to rate changes in audience engagement at the fine-grained level within talks, also providing evidence of the importance of these areas of emphasis as perceived by the audience at the time of the presentation.

- Research Question 3 (RQ-3-1): Can visual and acoustic stimuli on the part of the speaker be utilised to discover areas of special emphasis being provided by the speaker to indicate important parts of their presentation?
    - Secondary RQ (RQ-3-2): If we can detect spoken emphasis, is there a relationship between speaker ratings and emphasised speech, and between audience engagement and emphasised speech?

Another dimension of presentations is their potential to be comprehended. The concept of audience comprehension has to date not been studied in the domain of academic presentations or lectures. However, other studies have been performed on spoken language comprehension. (Tanenhaus et al., 1995) used eye-tracking to study the effects of relevant visual context on the mental processes that accompany spoken language comprehension. The effect of the design of presentation slides on audience comprehension was studied by (Garner and Alley, 2013). In this work they tested the effect on audience comprehension of slides that adhered to six multimedia principles versus slides that followed commonly practised defaults in Microsoft PowerPoint. Language comprehension in children was studied by (Haake et al., 2014), who observed that a faster speech rate had a negative effect on children's language comprehension while a slower speech rate had a positive effect. A study of comprehension of non-native speakers was performed by (Lev-Ari, 2014), who used eye-tracking to show that when following instructions from non-native speakers, listeners make more contextually-induced interpretations. The author also suggests that those with relatively strong working memory also tend to increase their reliance on context to anticipate the speaker's upcoming reference and are less likely to notice lexical errors in non-native speaker's speech.

- Research Question 4 (RQ-4-1): Can we utilise visual and acoustic stimuli to train a classifier to automatically identify levels of comprehension among the audience to a presentation?
  - Secondary RQ (RQ-4-2): If we can classify levels of audience comprehension, is there a relationship between audience engagement and audience comprehension?

Current summarisation of academic presentations relies on the words spoken. In this thesis we explore the inclusion of automatically classified paralinguistic features, namely audience engagement levels, areas of intentionally or unintentional emphasised speech, and areas of high potential audience comprehension of the speaker to enhance the summarisation process. A system for analysing and annotating video sequences of technical talks was presented in (Ju et al., 1998). This used a robust motion estimation technique to detect key-frames and segment the video into sub-sequences containing a single slide. (He et al., 1999) used prosodic information from the audio stream to identify speaker emphasis during presentations, in addition to pause information to avoid selecting segments which start mid-phrase. They also garnered information from slide transition points to indicate the introduction of a new topic or sub-topic. They developed three summarisation algorithms for slide transition based summary, pitch activity based summary and summary based on slide transitions, pitch activity and user-access information. The approach outlined in this thesis goes beyond this by including areas of emphasis, high audience engagement and potential comprehensibility in generated summaries.

An enhanced digital video browser was developed and evaluated in (Li et al., 2000). This evaluated the effectiveness of the following enhanced browser controls - Time Compression (TC), Pause Removal (PR), Table of Contents (TOC), Shot Boundary (SB), Timeline markers and jump controls. We use this idea of enhanced browser controls for further evaluation of generated summaries by comparing automatically generated summaries to the use of such enhanced digital video browsers for gaining a quick overview of the presentation. (Joho et al., 2009) captured and analysed the user's facial expressions for the generation of perception-based summaries which exploit the viewer's affective state, perceived excitement and attention. Perception-based approaches are designed to overcome the semantic gap problem in summarisation, the rich meaning expected by a user versus the shallowness of automatically extracted content descriptions, by finding affective scenes in video. Our approach attempts to bridge this semantic gap by detecting and including in summaries areas of high engagement and comprehensibility for an audience. A set of tools for creating video digests of informational video is described in (Pavel et al., 2014). Informal evaluation suggests these tools make it easier for authors of informational talks to create video digests. (Chen and Withgott, 1992) studied the use of emphasis for automatic summarisation of a spoken discourse. Emphasised speech from one speaker was detected and summarisation excerpts were extracted with no noticeable differences with human extracted summarisation excerpts. A HMM was used for training the emphasis detection model. Our approach extends this work by including highly engaging and comprehensible areas of presentation in addition to detected areas of emphasis.

- Research Question 5 (RQ-5): Can areas of special emphasis provided by the speaker, combined with detected areas of high audience engagement and high levels of audience comprehension, be used as a component in the effective summarisation of academic presentations?

1.5 Structure

The remainder of this thesis is structured as follows:

Chapter 2 provides a literature survey of related work. We provide an overview of current audio and visual feature extraction techniques. These techniques are reviewed to discover which of these are most useful for work of the nature examined in this thesis. Following this a review of previous work in the area of video summarisation is provided, including the two basic approaches of keyframe extraction and video skims. A short introduction to multimodal signal fusion is then presented, including common techniques and challenges for the fusion of multimodal data signals. Previous work is also introduced on the classification of the concepts of audience engagement, 'good' speaking techniques, speaker emphasis and audience comprehension.

In **Chapter 3**, an introduction is provided to the multimodal dataset we developed for use in this research. This is a collection of audio-visual recordings of academic presentations from an international conference. The conference, venue, number of presentations, number of presenters, general layout of conference and recording details are introduced. Details are also provided of the human annotation performed over this dataset for each experiment performed.

Chapter 4 describes the experiments performed and results achieved for the classification of areas of 'good' public speaking techniques. We then explain how this can be used to classify areas of audience engagement by utilising audio and visual features provided by the presenter. In the second part of this chapter, we describe an investigation into the identification of emphasised content during presentations and explore potential correlations between areas of intentionally or unintentionally emphasised speech and fine-grained audience engagement levels.

In **Chapter 5**, we report on the prediction of comprehension levels among the audience to academic presentations. This chapter describes the experiments performed and results achieved. We experiment with different fusion techniques and demonstrate how the decision-level fusion (post-fusion) of selected pre-trained classifiers can improve classification accuracy. We also demonstrate the problem of predicting a presenter's potential to be comprehended by an audience. This chapter experiments with the use of feature-level fusion and decision-level fusion techniques for the fusion of multiple modalities.

Chapter 6 reports on the development of automatically generated video summaries based on the methods introduced in Chapters 4 and 5. Techniques used in this summarisation strategy and reasoning for these are described. An evaluation of summaries is performed, including eye-tracking evaluation, and the comparison of automatically generated presentation summaries with video skim techniques previously found by (Li et al., 2000) to aid users gain a quick overview of presentations and lectures for further evaluation. Evaluations are also performed to compare usage of a subset of all available features.

Chapter 7 provides final answers to the research questions addressed in this thesis in addition to conclusions from this research and suggests some future research directions which could further advance this work.

Appendix A, B & C We provide a list of all publications emanating form this work. Following this, we provide additional results to the experiments conducted in Chapter 6. Additional technical information on the tools used during the course of this research is also provided.

Chapter 2

Background Review

In this chapter we provide a review of previous work related to the topics investigated in this thesis. We begin by reviewing work in the area of non-verbal communication. This leads into our review of work directly related to the topics investigated in this thesis. Following this, we review work in the area of visual and audio feature extraction. We then review work in the area of video summarisation and take a look at the two main approaches to summarisation of video: keyframe extraction and video skims. Following this, we undertake a review of the main challenges involved in multimodal signal fusion.

2.1 Non-Verbal Communication

Non-verbal communication includes any non-speech communicative acts intended to portray some meaning. Typical non-verbal communications include facial expressions, posture, gesturing, head movements, laughs etc. Non-verbal expressions can be seen as social displays as well as emotional expressions. Smiling is one of the most common nonverbal signals used in communication among humans. (Kraut and Johnston, 1979) showed that smiling occurred primarily in the presence of a receiver rather than in non-social but happy circumstances. This indicates that many non-verbal signals are primarily social displays, despite being obvious emotional expressions.

> The broad definition of non-verbal communication includes any kind of non-verbal messages or non-verbal signs proper to informative processes. A narrow definition restricts it to non-linguistic phenomena that are inter-related, often in an intricate way, with verbal language and can be found in interactive or communicative processes.
>
> (Payrató 2009:164)

(Dynel, 2011) reviewed the work of (Grice, 1991) and others on non-verbal communications from the narrow definition pertaining to human communication via body language. One of Grice's key tenets is that intentionality underlies non-natural meaning as opposed to natural meaning. He uses the latter term in reference to any stimulus which conveys some information.

Some researchers e.g. (Morris and Morris, 1977; Gibbs, 1999; Payrató, 2009) distinguish between several constituents of non-verbal communication:

- polemics (physical distance in communication: intimate, social, personal, public)
- chronemics (communication time)
- haptics (touch in communication)
- kinesics (movement, posture, gesture, facial expression, gaze)
- para-language (non-verbal speech aspects, prosody)
- olfactics(smell)
- oculesics (eye movement)
- physical appearance (clothing, hairstyle etc.)
- artefacts (manipulative objects).

The two most frequently discussed categories of non-verbal communication are para-language and kinesics.

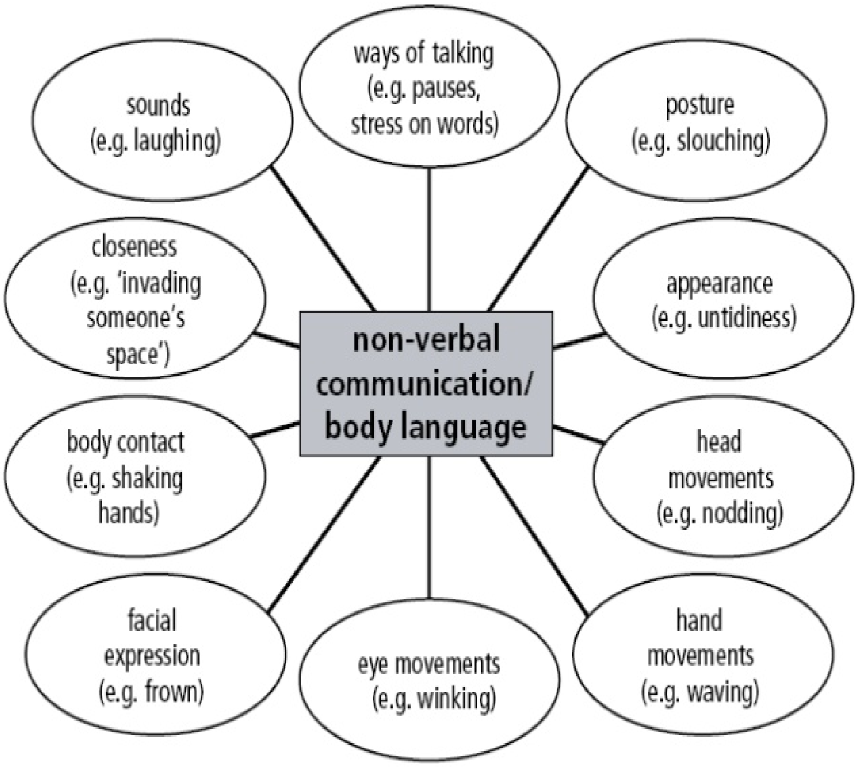
- Para-language - the audio aspects such as pitch, intensity and fluency of speech etc., are particularly important aspects for speaker ratings, engagement, emphasis of speech and the comprehensibility of a presentation.

- Kinesics - the movement, gesticulation, gaze and posture of a speaker is particularly important to determining the engagement of a presentation, emphasised parts of speech, and may also play a key role in the comprehensibility of a presentation.

Gestures, a common form of non-verbal communication, are defined as body activities (non facial) which do not include postural, spatial or orientational actions. (Kendon, 2004) endorses a continuum of gestures ranging from gesticulation, language-like gestures, pantomimes, emblems and sign language. This continuum shows a decline in importance of accompanying speech. In turn, the continuum shows an increase in gestures' likeness to utterances. As one of the most common forms of nonverbal communication, we need to estimate a speaker's gesticulation to be able to make inferences as to the overall interestingness / comprehensibility of a presentation.

Gesticulation is a seemingly spontaneous, unrehearsed body activity which co-occurs with speech, and is perceived as forming an intimate part of the total utterance (McNeill, 1992). Gesticulatory movement obligatorily accompany speech and shows a “lack of language-defining properties, idiosyncratic form-meaning pairings, and a precise synchronization of meaning presentations in gestures with co-expressive speech segments” (McNeill, 2000). Gesticulation can include baton or beat gestures.

Language like gestures are used in place of words and are syntactically integrated into an utterance but are not used by convention. Pantomimes, in turn, occur independently of speech and amount to sequential demonstrations. These are special gestures which rely on working memory. A common example would be a person



non-verbal communication/ body language
sounds (e.g. laughing)
ways of talking (e.g. pauses, stress on words)
posture (e.g. slouching)
closeness (e.g. 'invading someone's space')
appearance (e.g. untidiness)
body contact (e.g. shaking hands)
head movements (e.g. nodding)
facial expression (e.g. frown)
eye movements (e.g. winking)
hand movements (e.g. waving)

Figure 2.1: Types of non-verbal communication, from (Non, 2014)

using their finger to mimic a toothbrush while acting out brushing their teeth. Both language like gestures and pantomimes may entail the use of objects. A distinct subtype is distinguished from language like gestures and pantomimes, namely deictic, or pointing gestures, which index referents.

Emblems or quotable gestures are coded gestures emerging through conventionalisation processes. Emblems may derive from gesticulation or illustrators which have undergone conventionalisation. An example could be a presenter making quotation emblems with their fingers during a presentation whilst saying "good", which would more accurately mean 'commonly perceived as good'. Emblems are used in place of their verbal counterparts and enjoy conventional meanings. Figure 2.1 above summarises different types of non-verbal communication.

2.2 Background and Related Work

This section reviews prior work directly relating to the topics investigated in this thesis. This will look at what has already been studied, what has yet to be investigated, and how the prior work informs our decision making. We look at previous work in the area of public speaking skills, engagement detection, emphasis detection and audience comprehension.

2.2.1 Rating of Public Speaking and Presentation Skills

In this section we review existing research in public speaking and presentation skills. We first look at investigations examining what constitutes good speaking techniques. Previous work has studied what people popularly perceive to be 'good' public speaking skills, extracted features from the audio track and evaluated the correlation with these features and human judgements (Strangert and Gustafson, 2008; Chen et al., 2014; Liscombe et al., 2003). We also look at previous work in the area of engagement detection, most commonly performed in a multimodal meeting context. Separate studies have looked at engagement detection, most commonly in meetings and usually through the modalities of eye-gaze and facial expressions.

(Strangert and Gustafson, 2008) looked at qualities of a good public speaker in the context of political speech. Human annotators listened to clips of political speeches from the Swedish parliament and rated the qualities of each speaker according to a number of statements about that speaker. Acoustic measurements were taken to study the most important aspects which showed F0 to correlate with positive statements about speakers. F0 is a formant in speech, it has a one-to-one relationship with pitch, hence these two terms are often conflated. F0 dynamics were shown to influence the impression of charisma / good speaking across widely different languages. F0 features, particularly a wide F0 range and high peaked focussed words were found to give high ratings of a 'good speaker'. Results from this indicate that F0 dynamics > fluency > speech rate when it comes to perceptual weight.

An assessment of good public speaking skills was performed by (Chen et al., 2014) who extracted multimodal cues and evaluated the linear correlation with human holistic scores - overall grades applied to speaking skills. In this work they found that simple multimodal features of the speech content, speech delivery and non-verbal behaviours together can predict human scores on presentation performance with significant accuracy. Features of the content of speech were extracted using a syntactic complexity analyser tool on speech transcripts (Lu, 2010). This tool counts the frequencies of nine types of syntactic structures e.g., verb phrases, clauses etc., and computes fourteen syntactic complexity values such as mean length of clause.

(Liscombe et al., 2003) trained a classifier on emotional speech using acoustic features. In this work they used an emotional speech corpus compiled using eight professional actors. Subjects listened to emotional speech samples and expressed a judgement on the emotion being expressed. The automatically extracted features include min F0, max F0, mean F0, F0 range, F0 standard deviation, minimum amplitude, maximum amplitude, mean amplitude, range amplitude, standard deviation amplitude, ratio of voiced samples to total segments and mean syllable length, spectral tilt, contour and type of nuclear accent. They found that F0, RMS, and speaking rate are good at distinguishing emotions on the grounds of activation. However, this study also found that spectral tilt and type of phrase accent and boundary tone may be useful in discriminating between the valency of emotions, with *friendly*, *happy* and *encouraging* falling into one category and *angry* and *frustrated* into another.

A multimodal virtual audience platform was developed by (Batrinca et al., 2013) for public speaking training. In this work they use professional public speakers invited from toastmasters to rate the qualities of presentations using audio and visual stimuli. They found several expert estimates of non-verbal behaviour to be significantly correlated with an overall assessment of a presenters performance. Assessed behaviours of presenter's in this work include flow of speech, clear intonation, interrupted speech, speaks too quietly, vocal variety, paces too much, gestures to emphasise, gestures too much, gazes at audience and avoids audience.

Toastmasters International have published a practical guide to becoming a better speaker (International, 2015). In this they emphasise the importance of body language. According to this guide, not only does body language communicate confidence and power, but enhances your believability and emphasises points you are making. Body language is expressed in stance, movement, gestures, facial expressions and eye-contact. They also emphasise the importance of vocal variety, such as volume (Intensity), pitch and speech rate. A speech rate which is too fast will cause the audience to not be able to keep up, which a speech rate which is too slow will cause the audience to lose interest. Variety in both loudness and pitch is also important for maintaining the audiences attention.

A speaker's stance, facial expressions and eye-contact are outside of the scope of this thesis due to the difficulty in extracting these features. In this work we focus on the extraction of movement, gestures, volume, pitch and speech rate.

2.2.2 Detection of Engagement

In this section we look at previous work on engagement detection and the relationship between group-level and individual-level engagement.

(Oertel et al., 2011) aim to better understand the dynamic changes in human interaction in order to add social information to speech technologies. They studied automatic detection of involvement in conversation through measuring body movements and voice features. Involvement was annotated on a scale of 1-10, though all annotations were labelled from 4-9. This study found a clear linear relationship between their perceptual measure of involvement, level and span of the voice as well as intensity. These results suggest that involvement seems to be a scalar rather than a binary phenomenon.

Conversational engagement in multi-party video conversation was studied by (Bednarik et al., 2012), who focussed on the estimation of conversational engagement from gaze signals. In this work they used 6 levels of engagement (no interest, following, responding, conversing, influencing, managing) annotated over 15 second intervals using two annotators. In this they focused on the individual within the group rather than the group as a whole. In this work they found that gaze behaviour differs during distinct levels of conversational engagement. They built an SVM classifier using gaze-based features which correctly predicts engagement to 74%. They found that with increasing activity and engagement in discussion, the mean fixation duration drops and there is also a decrease in the number of fixations.

Group involvement was contrasted with individual engagement by (Bonin et al., 2012), who annotated conversational engagement at an individual level and at the group level, and relate individual involvement with group level engagement. No prior examples of engagement levels were provided to annotators so as not to influence their interpretations. Annotators watched the video once per speaker and marked any changes of involvement at either the individual or group level. From their annotations it became clear to analysts that certain participants seemed to be more important in the perception of group involvement with respect to others.

This work was extended by (Oertel and Salvi, 2013) who based their analysis on an increased 8-party conversation and proposed a new set of features using eye gaze to relate group involvement to individual engagement. In this work participants rated their own engagement levels. They found that it was possible to estimate individual engagement and group involvement by analysing the participants eye gaze patterns. They built a classifier able to distinguish between four classes, low, high, lead and organising, of group involvement with an accuracy of 71%.

(Grafsgaard et al., 2014) analysed the additive effect of multimodal features for predicting engagement, frustration and learning in the study of introductory computer programming tutoring in which tutors communicated with students through a text-based interface. They found large improvements from the unimodal (dialogue, nonverbal, task) set to the bimodal (dialogue * nonverbal, dialogue * task, nonverbal * task) set and from the bimodal set to the trimodal (bimodal union, dialogue * nonverbal * task) set. The complete trimodal set of features was found to be the most predictive.

Group engagement detection in multimodal meetings was performed by (Gatica-Perez et al., 2005). In this work annotators were asked to label 15 second intervals from 4-party meetings according to 5 levels of engagement. Audio features, including energy, pitch and speaking rate, and visual features, including global person motion, eccentricity and pose, were extracted from the multimodal corpus. Hidden Markov models (HMM's), a statistical Markov model in which the system being modelled is assumed to be a Markov process with hidden states, were used to detect segments of high and neutral group interest levels. In this work they found that audio only modalities performed better than visual only modalities. (McCowan et al., 2005) use HMM's to examine the relationship between individual and group level engagement. In this they made use of visual and acoustic features with no eye gaze information. They found that it is important to model the correlation between the behaviour of different participants and that there is evidence of asynchrony between participants acting within the group actions.

The automatic recognition of student engagement was studied by (Whitehill et al., 2014) from analyses of facial expressions. Undergraduate students had their faces recorded while using a cognitive skills training software application installed on an iPhone. For evaluation they used undergraduate and graduate students of computer science and psychology to rate the engagement level of each student using the software application based on their recorded facial expressions. They found that observers rely on head pose and elementary facial actions like brow raise, eye closure and upper lip raise to make judgements on student engagement.

Video analysis was performed by (Jang et al., 2014) to build a classifier for detection of engagement levels in children while using a robot based math game. Each video was annotated by three coders, and engagement was encoded in two states of engaged and not engaged. Sliding time windows of length 1 second to 6 seconds were evaluated, over several classification models (C4.5 decision tree classifier, Naive Bayes classifier, multi-layer perceptron and random forest). In their evaluation they found that a C4.5 decision tree classifier with time window length of 1 second performed best.

From the above cited previous work on speaking techniques and on engagement / involvement detection we can see that audio-visual features have been shown to be effective for building a classifier to automatically predict speaking traits. Engagement detection has not been studied as yet in the context of academic presentations. In this work we seek to address this gap in existing work by constructing a classifier to predict speaker ratings for each presenter. Further, we explore the development of a classifier intended to predict audience engagement by extracting audio-visual features from the presenter in addition to the audience, and fusing with automatically classified ratings of the speaker.

2.2.3 Emphasis Detection

In this section we review existing research related to our work on emphasis detection.

The use of emphasis for automatic summarisation of a spoken discourse was studied in (Chen and Withgott, 1992). Emphasised speech from one speaker was detected and summarisation excerpts extracted with no noticeable differences from a baseline of human extracted summarisation excerpts. Two sources of data were used for this investigation, a 27 minute long videotaped interview between two primary speakers and the second was a set of phrases extracted from telephone conversations, developed under a DARPA program. The emphasis detector was based on a Hidden Markov Model in which a separate model was created for each of the 3 levels of emphatic speech, unemphatic speech and background speakers.

Pitch based emphasis detection for automatic segmentation of speech recordings was explored in (Arons, 1994). Their initial investigation was based on recordings of talkers introducing themselves and presenting a 10 to 15 minute summary of their background and interests. A dialogue was transcribed and manually annotated with paragraph breaks and emphasised regions by a linguist. Based on preliminary analysis and investigations, a pitch threshold of the top 1% of pitch values was chosen, in which case speech segments with pitch values exceeding this threshold were classified as emphasised speech. This threshold was selected as a practical starting point, which could be varied to find more or less emphasised regions. From this, the pitch based segmentation technique could be used to summarise the speech recordings into the most important speech segments. Three monologues were segmented using this technique, and were highly correlated with topic introductions, emphasised phrases and paragraph boundaries in the transcript annotated by the linguist.

(He et al., 1999) attempted to summarise audio-visual presentations by exploiting information in the audio signal, knowledge of slide transition points, and information about access patterns of previous users. In this work they found that overlap between pitch-based segments and author-generated segments performed no better than would be achieved by random chance. From this, it appears that audio-visual presentations may be less susceptible to pitch based emphasis analysis than the audio stream only, or that spoken emphasis did not truly correspond to semantically important material.

Following on from this work, (Kennedy and Ellis, 2003) studied emphasis detection for characterisation of meeting recordings. In this work they had 5 human annotators label 22 minutes of audio from the International Computer Science Institute (ICSI) meeting corpus (Morgan et al., 2001). Annotators were given both an audio recording and a transcript from the meeting in which the annotators listened to the audio recording while working their way through the transcript and marking each utterance as emphasised or not. The authors extracted pitch and aperiodicity of each frame and calculated the mean and standard deviation for each speaker. In cases where 4 or more human annotators agree on emphasis, accuracy rates of 92% were achieved. In addition, the utterances found to be the most emphasised were indicated to be a good summarisation of the meeting recording by human annotators.

To the best of our knowledge, the detection of regions of speech emphasis has not previously been performed in an audio-visual context. As described above, pitch thresholds in the top 1 percentile were reliably found to have been emphasised by human assessors. However, (He et al., 1999) indicate that emphasis in audio-visual recordings is indicated by more than just notable increases in pitch in the audio stream, though findings by (Kennedy and Ellis, 2003) concluded that utterances found to be most emphasised represented a good summary of audio-only recordings of meetings. As this previous work shows a possible difference between emphasis in the audio-only stream and the audio-visual stream, in our study we investigate use of audio-visual features to detect emphasis in academic presentations.

2.2.4 Audience Comprehension

In this section we review existing research in audience comprehension. In preceding sections, we have already looked at investigations into what constitutes 'good' public speaking techniques and engagement. We now look at the small amount of previous work in the area of audience comprehension. We conclude this section by outlining how our work seeks to extend this earlier research.

Eye-tracking was used in (Tanenhaus et al., 1995) to study the effects of relevant visual context on the mental processes that accompany spoken language comprehension. We consider that this could also relate to an audience as they garner the visual context from presentation slides whilst listening to the presentation speech. They show that in natural contexts, people seek to establish visual reference with respect to their behavioural goals from the earliest moments of linguistic processing. Moreover, referentially relevant non-linguistic information immediately affects the manner in which linguistic input is initially structured.

How the design of presentation slides affects audience comprehension was studied in (Garner and Alley, 2013). In this work they tested the effect of slides which adhered to six multimedia principles on audience comprehension versus slides which instead followed commonly practised defaults in Microsoft PowerPoint. The slides which adhered to the six multimedia principles followed the assertion-evidence approach. This is an approach to creating and delivering presentations by building a talk on messages, not topics, to support these messages with visual evidence, not bullet lists, and to explain this evidence by fashioning words on the spot. Participants were required to relate the process of Magnetic Resonance Imaging (MRI) in an essay. Essay responses from 110 engineering students revealed superior comprehension levels and fewer misconceptions for the assertion-evidence group in addition to a lower perceived cognitive load. Cognitive load was self-rated by students as their perceived mental effort on a 7-point scale.

Language comprehension in children was studied by (Haake et al., 2014), who observed that a faster speech rate had a negative affect on children's language comprehension using the Test for Reception of Grammar, version 2 (TROC 2) (Bishop, 2009), while a slower speech rate had a positive affect. However, for more difficult test items in TROC 2, the benefits of a slower speech rate were only pronounced for the children who had scored better on a working memory test.

A study of comprehension of non-native speakers was performed by (Lev-Ari, 2014) who tracked eye movements of participants to show that when following instructions from non-native speakers, listeners make more contextually-induced interpretations, increasing their reliance on context rather than depending on the speakers language alone. The author also suggests that those with a relatively strong working memory also tend to increase their reliance on context to anticipate the speaker's upcoming reference and are less likely to notice lexical errors in the non-native's speech, indicating that they take less information from the non-native speaker's language.

Our current work aims to integrate prior work on speaking techniques and audience reactions and engagement to the concept of potential audience comprehension. We do this by studying the relationship between extracted audio-visual features and human-holistic scores of comprehension levels during academic presentations, and subsequently training a classifier to identify comprehension levels on these extracted features.

We study pre-fusion and post-fusion techniques to discover the best feature-fusion strategies for work of this nature. We also calculate correlations between comprehension and audience engagement in order to study potential relationships between these two concepts.

To the best of our knowledge, no current work exists which attempts to classify potential audience comprehension of audio-visual material or to relate this concept to audience engagement.

2.2.5 Academic Presentation Summarisation

In this section we look at related work on the summarisation and skimming of academic presentations.

The use of motion estimation techniques for analysing and annotating video recordings of technical talks was investigated in (Ju et al., 1998). They used a robust motion estimation technique to detect key frames and segment the video into sequences containing a single slide. For image sequences corresponding to a particular slide, stabilisation is warping each of the images towards a reference image, taking into account the cumulative motion estimated between each of the frames in the sequence. Potential gestures are tracked using active contours, found by computing the absolute difference between the key frame and images in the warped sequence. By successfully recognising all pointing gestures, presentations can be fully annotated per slide. This automatic video analysis system helped users to access presentation videos intelligently by providing access using specific slides and gestures.

Prosodic information from the audio stream to identify speaker emphasis during presentations was used in (He et al., 1999), in addition to pause information to avoid selecting segments for summaries which start mid-phrase. They also garnered information from slide transition points to indicate the introduction of a new topic or sub-topic. They developed three summary algorithms: a slide transition based summary, a pitch activity based summary and a summary based on slide, pitch and user-access information. They used surveys for evaluation and found that computer generated summaries were rated poorly on coherence, in which participants complained that summaries jumped topics. No significant difference between users' preferences for the three methods was found, leading to the conclusion that the simpler methods may be preferable. They also found that audio-visual presentations were less susceptible to pitch-based emphasis analysis than the audio-only stream, meaning emphasise is more easily analysed from pitch in the audio-only stream. This is the first work attempting to summarise presentation video by speaker emphasis. In this work they found that speaker emphasis was not sufficient to generate effective summaries.

(Li et al., 2000) developed and evaluated an enhanced digital video browser for browsing through different categories of digital video. The categories evaluated were classroom, conference, sports, shows, news and travel. They evaluated the effectiveness of the following enhanced browser controls - Time Compression (TC), Pause Removal (PR), Table of Contents (TOC), Shot Boundary (SB), Timeline markers and jump controls. For browsing of conference presentations, TC and PR were found to be the most effective tools for improved video browsing with scores of 6.9 and 6.5 out of 7 respectively. This enhanced digital video browser can give us an additional comparison method to evaluate the effectiveness of our automatically generated presentation summaries, to complement analysis using eye-tracking methods.

(Joho et al., 2009) captured and analysed user's facial expressions for the generation of perception based summaries which exploit the viewer's affective state, perceived excitement and attention. For their work they used the piecewise Bezier volume deformation tracker (Tao and Huang, 1998). Perception based approaches are designed to overcome the semantic gap problem in summarisation by finding affective scenes in video. They find it unlikely that a single summary could be generally seen as highlighting the key features of a video by all viewers. Results suggest that there were at least two or three distinguished parts of videos that can be seen as the highlight by different viewers.

(Pavel et al., 2014) created a set of tools for creating video digests of informational videos. These tools include text summarisation, chapter and section segmentation, and had a video digest authoring interface. Informal evaluation suggested that these tools make it easier for authors of informational talks to create video digests. They also found that crowdsourced experiments suggest that video digests afford browsing and skimming better than alternative video presentation techniques.

Summarisation of academic presentations using high-level paralinguistic features has yet to be studied to the best of our knowledge. We look to summarise presentations by incorporating the parts found to be most engaging, emphasised and comprehensible for the audience. We contrast and combine these with summaries built using automatic speech transcripts and keywords of the presentation. The focus of this work is to develop automatic summaries of academic presentations which are as engaging and comprehensible as possible by using the most engaging and comprehensible parts while using important keywords to ensure that summaries are not only more engaging than the original presentations but also maintain their coherence. We also use the work from (Li et al., 2000), who found an enhanced digital video browser, with pause removal and the ability to view presentations at up to 2.5 times normal speed, to be very effective for gaining a quick overview of presentations, as guidance for an additional novel evaluation strategy, by crowdsourcing questionnaires and comparing to automatically generated summaries for effectiveness and ease of use.

2.3 Visual Feature Extraction

Visual features can be extracted automatically from the visual stream. They can be used to assist in browsing, summarisation and retrieval of multimedia content by searching for and tracking those specific features. These can include scene types, textures, colours, shapes as well as objects such as human faces. Extraction can also be performed on low-level local features for corner and edge detection and foreground and background segmentation, motion detection and object tracking. These can be performed frame-by-frame in a video stream to track movements of objects through the visual stream. Table 2.1 lists different types of visual feature detectors.

There are many types of visual salience (noticeable, of interest) classifiers which can be used for extraction of features from the video stream. Many of these are based on local image features, regions which display a certain amount of non-uniformity in intensity values. Examples include edge or corner detectors which use intensity to identify edges or corners in pictures. These features can be used to find correspondences between sets of images and are more discriminative than low-level features. Many such curvature feature detectors are based on edges, colour or texture.

Interest point algorithms, which identify potential points of interest in images and use these for edge or corner detection or pattern recognition, such as Scale-Invariant Feature Transform (SIFT), should ideally be invariant to scale, translation and rotation. These should also be partly invariant to small affine changes or changes in illumination. The most important part of a local image feature detection algorithm is repeatability, the ability of the algorithm to detect the same interest points in different images.

The following are examples of visual feature detectors which are typically used for visual feature detection, which could potentially be useful for video summarisation work:

Table 2.1: Visual Feature Detectors

Type	**Name**
Low-Level	Edge Detection
	Corner Detection
	Blob Detection
	Ridge Detection
	Scale-Invariant Feature Transform
Curvature	Edge Detection
	Changing Intensity
	Autocorrelation
Image Motion	Motion Detection
	Area Based
	Differential Approach
	Optical Flow
Shape-Based	Thresholding
	Blob Extraction
	Template Matching
	Hough Transform

Given that for videos in the domain of speaker presentations, the main visual stream is of a single speaker presenting in front of an audience, and that the secondary stream is of the audience to this presentation, we deem the most useful visual feature detectors to be Optical Flow for advanced motion estimation to account for head movement and gesticulation as well as audience movement, and facial detection in order to get the position of the speakers face whilst speaking. We consider the usefulness of other visual detectors in this stream to be rather limited given that much of the information for these videos is actually contained in the spoken audio stream.

The detection of salient regions of images is now relatively mature and several algorithms exist to detect salient regions of images. One of the first of these to emerge is the Harris Corner Detector (Harris and Stephens, 1988). This algorithm detects points on an image that are located on corners or vertical edges. Edge hysteresis is performed, enhancing the continuity of edges. Further processing deletes edge spurs and short isolated edges, and bridges short breaks in edges. This results in continuous thin edges that generally terminate in the corner regions. Edge terminators are then linked to the corner pixels residing in the corner regions to form a connected edge vertex graph. Figure 2.2 demonstrates the Harris Corner Detector detecting keypoints.

Another well established algorithm is the Scale Invariant Feature Transform (SIFT) (Lowe, 2004), which detects distinctive, salient regions, and for creating highly discriminative feature vectors for these interest points. The feature vectors can be used for reliable matching of visually similar interest points in different images. The first stage of computation searches over all scales and image locations. This identifies potential interest points that are invariant to scale and orientation. Next, at each candidate location, a detailed model is fitted to determine location and scale. Then, one or more orientations are assigned to each keypoint based on local image gradient directions. Local image gradients are measured at the selected scale in the region around each keypoint.

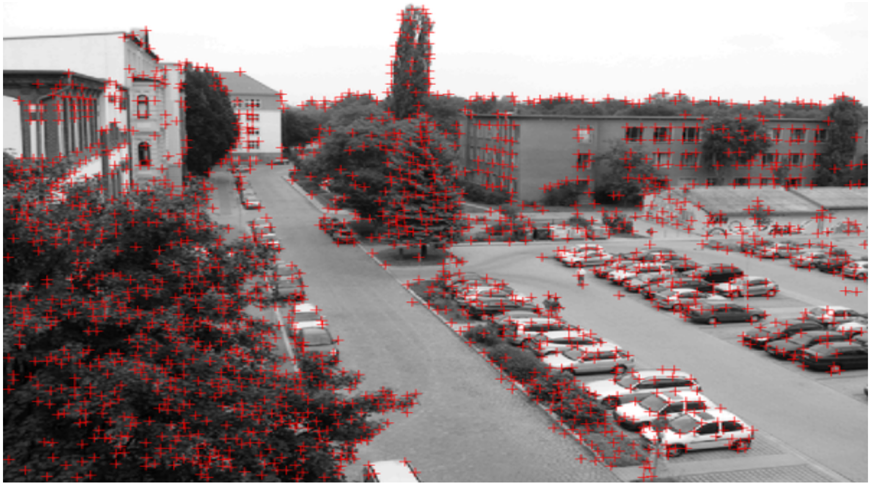


Figure 2.2: An example of Harris Corner Detector, keypoints detected at all major corners or edges, from (Har, 2017)

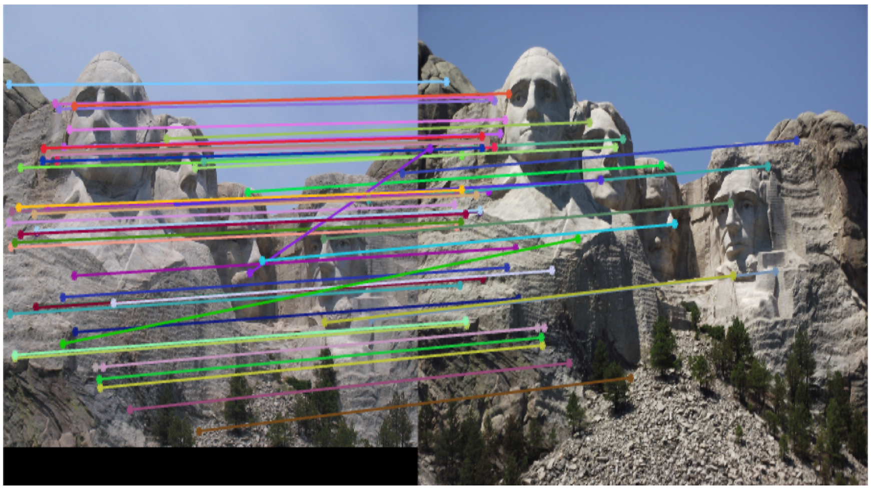


Figure 2.3: An example of Scale Invariant Feature Transform (SIFT), keypoints detect identical object in different images, from (Sif)

The Speeded Up Robust Features (SURF) algorithm is described in (Bay et al., 2008). This is based mainly on SIFT, and is now used in place of SIFT in many instances due to its superior speed. SURF is also claimed by the authors to be more robust against different image transformations (stretches or alternations of the image) than SIFT. Figure 2.3 shows an example of SIFT in which keypoints detect identical objects in different objects.

Another efficient alternative is Oriented FAST and Rotated BRIEF (ORB), described by (Rublee et al., 2011). ORB is an efficient alternative to SIFT and SURF, estimated to be two orders of magnitude faster than SIFT. Another advantage of ORB is that it is not patented and is thus free to use, whereas SURF and SIFT require license fees to use their original algorithms. ORB is a fusion of a Features from Accelerated Segment Test (FAST) keypoint detector, which is a method of choice for finding keypoints in real-time systems, and Binary Robust Independent Elementary Features (BRIEF) descriptor, a feature descriptor that uses binary tests between pixels in a smoothed image patch. BRIEF first uses FAST to find the keypoints, and then applies the Harris corner measure to find the top N points. ORB then steers BRIEF according to the orientation of the keypoints. ORB performs the efficient computation of oriented BRIEF features, then performs analysis of variance and correlation of oriented BRIEF features, and adds a learning method for de-correlating BRIEF features under rotational invariance which leads to improved performance in nearest-neighbour applications. Figure 2.4 shows an example orb descriptor detect identical objects in different images.

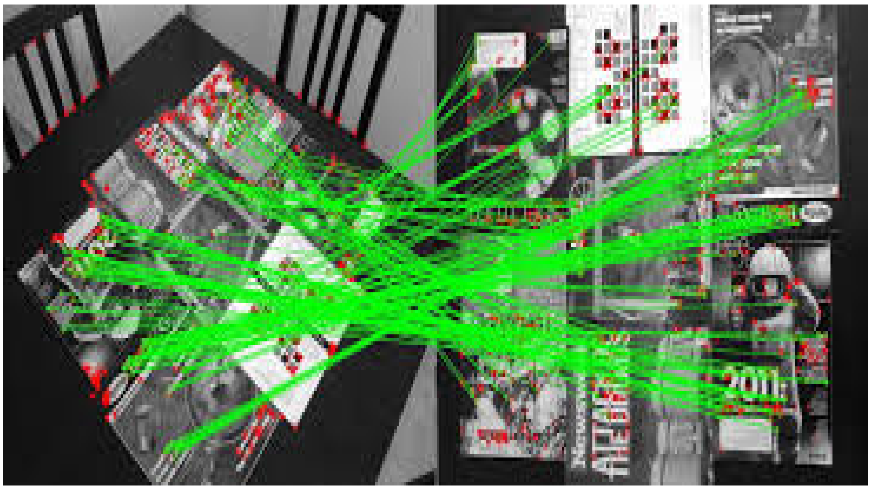


Figure 2.4: An example of ORB descriptor, similar to SIFT and SURF, keypoints detect identical objects in different images, from (Orb, 2013)

The above are examples of hand crafted features typically used for Computer Vision problems. For these features the data set does not have to be large, and you have the opportunity to explicitly appreciate features which largely contribute to the particular task. It may be challenging however to come up with such reasonable and consistent descriptors/features on some hard tasks. Newer approaches like convolutional neural networks do not need to be supplied with such hand crafted features as they are able to learn the features automatically from the data. However, Deep Learning requires much more training data than is required for hand-crafted features.

Due to the type of presentation summarised in this work, single scene conference presentations with the presenter(s) speaking at a podium, presenting slides where most of the information is in the audio stream, the hand crafted features described above are not of great use for the work in this thesis. Below, we describe the visual features which are of use for the work in this thesis, Face Detection and Optical Flow, which are used to recognise the audience as well as movement on the part of the speaker, and Optical Character Recognition, used for extracting information from the slides.

2.3.1 Robust Real-Time Face Detection

The work described in this thesis needs an efficient method to detect human faces in the visual stream. The accurate detection of faces allows us to estimate the orientation and position of the speaker throughout their presentation, it also allows us to estimate if the presentation is being given by more than one presenter. This also enables us to estimate the number of forward facing people in the audience, which could potentially be of benefit for estimating audience engagement.

Robust Real-Time Face Detection described in (Viola and Jones, 2004) is a popular solution to the problem of facial detection in the visual stream. There are three main contributions of Robust Real-Time Face Detection. The first was the introduction of a new image representation called an *integral image* which allows for very fast feature evaluation. This uses a set of features reminiscent of Haar Basis functions, which were previously introduced in (Papageorgiou et al., 1998). A Haar basis function is a sequence of rescaled 'square shaped' functions, which together form a wavelet basis. To rapidly compute these features at many scales, integral image representation is introduced. Here, the integral image is computed from an image using just a few operations per pixel.

The second contribution of (Viola and Jones, 2004) was a simple yet efficient classifier, built using a small number of important features from a large library of potential features using Adaboost (Freund and Schapire, 1995). Adaboost is a machine learning meta-algorithm which can be used in conjunction with many other types of learning algorithm to improve their performance. The output of these other learning algorithms (weak learners) is combined into a weighted sum that represents the final output of the boosted classifier. In order to ensure fast classification, the learning process must exclude a large number of available features, focussing on a small set of critical features, as the use of all features would make the learning process too cumbersome. Feature selection is achieved using Adaboost by constraining each weak classifier to depend on a single feature. This means that each stage of the boosting process which selects a new weak classifier can be viewed as feature selection.

The third and major contribution of the work by Viola & Jones, is a method for combining complex classifiers in a cascade structure which dramatically increases the speed of the detector by focussing attention on promising regions of the image. The notion behind this is that it is often possible to determine where in an image a face might occur. More complex processing is then only reserved for these promising regions.

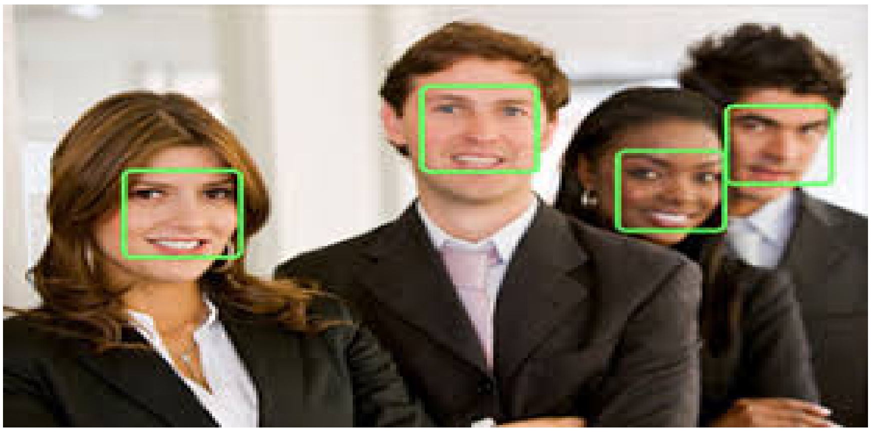


Figure 2.5: An example of Robust Real-Time Face Detection, from (Fac, 2015)

2.3.2 Lucas-Kanade method for Optical Flow

Optical Flow is a method for estimating the movement of interesting features, and could be very useful for the work described in this thesis due to its ability to estimate the direction of motion. This has the potential to give us more detailed information of gesticulation on the part of the speaker than simple motion detection. This can also give more detailed information on the movement of the audience, which may be of value for estimation of audience engagement.

Optical Flow (Lucas et al., 1981) is a simple algorithm for the estimation of movement of interesting features in successive images of a scene. The Lucas-Kanade algorithm makes some implicit assumptions:

- Two images are separated by a small time increment t, in such a way that objects have not displaced significantly as t moves to t+1.

- Images depict a natural scene containing textured objects exhibiting different intensity levels which change smoothly.

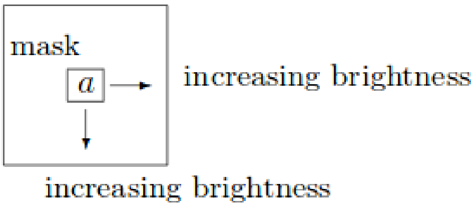


Figure 2.6: Optical Flow 1/2

The Lucas-Kanade algorithm does not use colour in any implicit way and does not scan the second image looking for a match for any given pixel. Instead, it works by trying to guess in which direction an object has moved so that local changes in intensity can be explained.

Let us watch a scene through a square hole in Figure 2.6. Intensity a, visible through the hole, is variable.

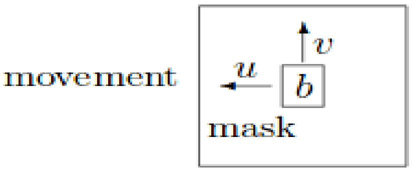


Figure 2.7: Optical Flow 2/2

The intensity of the pixel has increased to b in the next frame, Figure 2.7. We assume that the underlying object has been displaced to the left and up. These images are taken from Lucas-Kanade in a Nutshell by Prof. Raul Rojas[1].

The ability of this method to track the direction of movement means it is ideal for our goals, not only of tracking gesticulation on the part of the speaker, but also to track the very direction of the gesticulation, with movement in the opposite direction counting as more than continuous movement in the one direction. The same

[1] `http://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/Lucas-Kanade2.pdf`

reasoning applies for the tracking of movement among the audience to presentations, with movement in opposite directions counting as more than continuous movement in a single direction. Figure 2.8 shows an example of Optical Flow usage to indicate the direction and velocity of pedestrian movements.

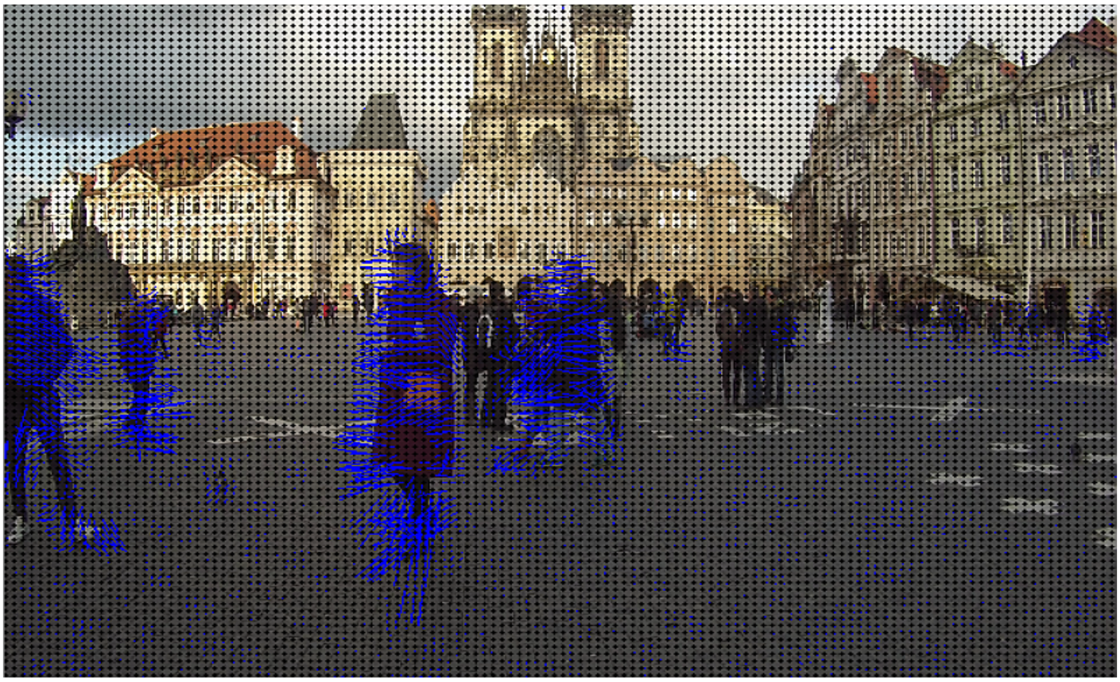


Figure 2.8: An example of Optical Flow: Clusters of blue pixels around pedestrians indicate their velocity and direction of movement, from (Opt, 2016).

2.3.3 Optical Character Recognition

Optical Character Recognition (OCR) is the recognition of printed or written text characters by a computer (Verma et al., 2016). It involves photo-scanning of the text, character-by-character, analysis of a scanned image and then translation of the character image into character codes, such as ASCII, which are commonly used in data processing. Figure 2.9 shows an example of OCR.

Optical Character Recognition is used in this thesis for the task of recognising characters from presentation slides. It is performed over images of each slide for each presentation, allowing us to gain an overview of the amount of text per slide, this enables inferences to be made from this which could aid the training of a classifier for identification of expected levels of audience comprehension based on the complexity of the slides being used in the presentation.

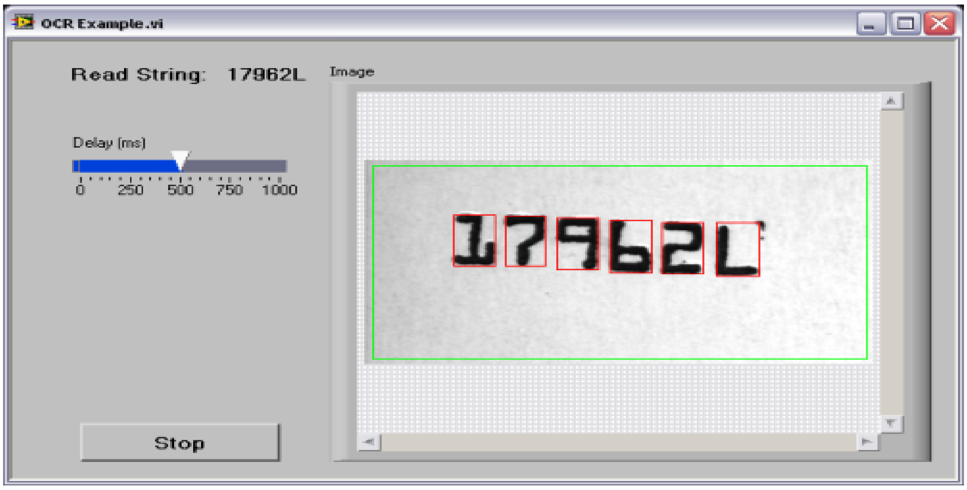


Figure 2.9: An example of Optical Character Recognition, from (Opt, 2015)

2.4 Audio Feature Analysis

Audio analysis refers to the extraction of information and meaning from the audio signal. This is a vital aspect of the work described in this thesis as audio features are the main information we have available for classification for the data set used in this thesis. The main audio features extracted and analysed in this work are pitch and intensity as well as speech formants. These features are very important for the overall interestingness, fluency and comprehensibility of a presentation, and thus form a vital part of the automatic audio-visual feature analysis performed for classification of the paralinguistic features in this thesis. Figure 2.10 demonstrates pitch and intensity profiles for male and female speakers.

2.4.1 Pitch

In speech, pitch is the relative lowness or highness of a tone as perceived by the ear (Takeuchi and Hulse, 1993; Olson, 1967). It depends on the number of vibrations per second produced by the vocal cords. It is the main acoustic correlate of tone and intonation. For men and women, the size difference of the vocal folds, reflecting differences in larynx size, will influence pitch range so that male adult voices are usually lower-pitched with larger folds than adult female voices. A speaker will use different patterns of pitch to attempt to convey different meanings to the listener.

Pitch is closely related to frequency but the two are not equivalent. Pitch is the auditory attribute of sound according to which sounds can be ordered on a scale form low to high. Since it is such a close proxy for frequency, it is almost entirely determined by how quickly the sound wave is making the air vibrate and has very little to do with intensity. High pitch means very rapid oscillation and low pitch means slower oscillation. Pitch depends to a lesser degree on the sound pressure level of the tone. The pitch of lower tones gets power as sound pressure increases.

2.4.2 Intensity

The intensity, often called the acoustic intensity, is perceived as the loudness of the sound (Int, 2017; Lou, 2016). The greater the intensity, the louder the sound is perceived to be. The lower the intensity, the quieter the sound is perceived to be. Intensity is controlled mainly by the force with which the air from the lungs is allowed to pass through the larynx. Voice pitch can remain constant whilst the loudness of that particular pitch can be varied. The frequency of vibration can be kept the same but to increase the amplitude of the vibration requires forcing more air through the larynx. The resulting speech sound wave produces greater motion in the air molecules. Vocal loudness will typically vary according to context, the speaker's mood and the content. Average speech intensity 65 dB SPL and has a range of 30 dB, any vowel has more power than any consonant.

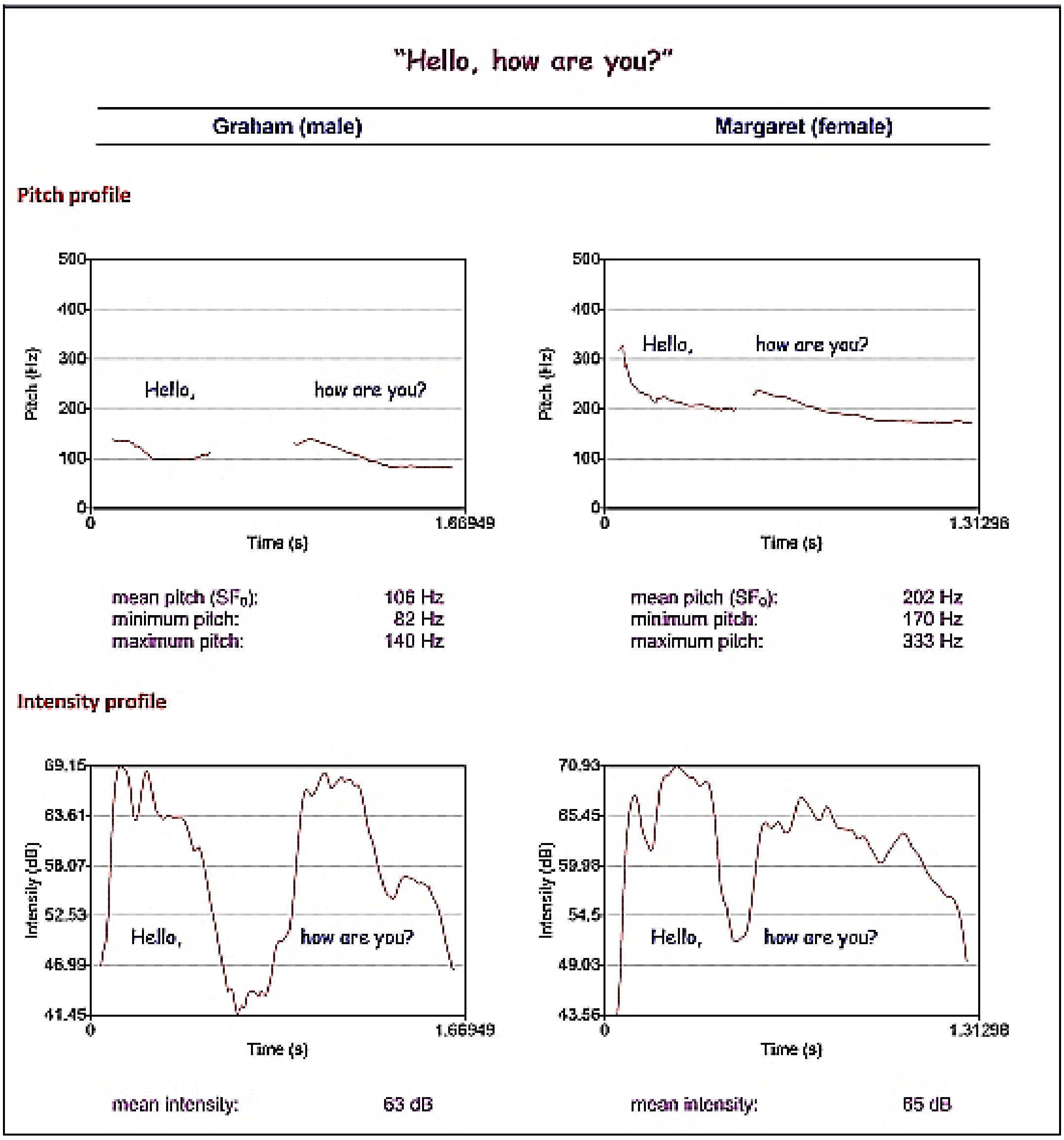


Figure 2.10: Pitch and Intensity Profiles for Male and Female Speakers. Image comes from (Lou, 2016).

2.4.3 Formants

Formants are frequency peaks which have a high degree of energy (Sundberg et al., 1977). They are particularly prominent in vowels. Each formant corresponds to a resonance in the vocal tract. During speech production, the source signal is filtered according to the morphology of the oral tract and the articulators. The formants are the peaks in the spectral envelope and are numbered F1, F2, F3 F4 etc. The first two formants contribute strongly to the differentiation of vowels from one another. It is possible to get a 1 to 1 relationship between given articulatory features and the value of the formants. Every formant is determined by the joint effect of different articulatory characteristics. For instance, a spoken *i* has a relatively low F1 value and a relatively high F2 value, the tongue being displaced to the front. However for a spoken *a*, it is the other way round. In back vowels, F1 and F2 are closer to one another, while in front vowels they are more distant. F3 is involved in the differentiation between rounded and unrounded vowels, *i* an *y* etc. In singing, F3 and F4 are very important formants, as they can be made much stronger in singing than in speaking. This makes the voice stand out over musical instruments. Figure 2.11 demonstrates the spectral envelope of an *i* sound as pronounced by a male speaker.

In the section of this work performing classification of comprehension levels, we make much use of speech formants. In addition to the formants already outlined in the above paragraph, we make use of features such as the differences between F2 - F1, F3 - F4 and F3 - (F2 - F1) etc.

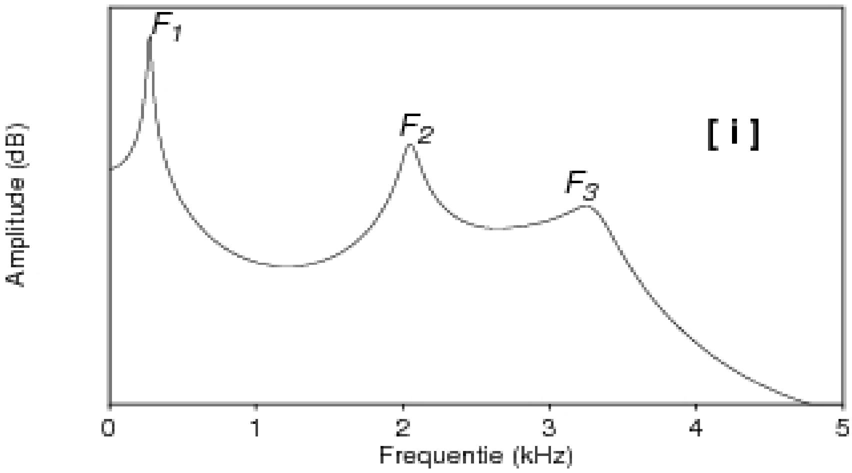


Figure 2.11: The spectral envelope of an *i* sound as pronounced by a male speaker. F1, F2 and F3 are the first three formants. This image is taken from (MAURO ANDREA VOICE STUDIO - CANTO AULAS)

2.4.4 MFCC's

In sound processing, Mel-frequency cepstrums (MFC) are representations of the short term power spectrum of a sound (Mermelstein, 1976). Mel-frequency cepstral coefficients (MFCC's) are coefficients that collectively make up an MFC. They are derived from a cepstral representation of the audio clip. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale. A mel scale is equivalent to a pyscho-acoustic scale in which we try to capture distances from low to high frequency. The following are the steps in computing MFCC's:

1. The first processing step is the computation of the frequency domain representation of the input signal. This is achieved by computing the Discrete Fourier Transform.

2. The second step is the computation of the mel-frequency spectrum. The spectrum is filtered with different band-pass filters and the power of each frequency band computed. The mimics the human ear because the human auditory system uses the power over a frequency band as signal for further processing.

3. The third step is to compute the logarithm of the signal to mimic the human perception of loudness.

4. The fourth step is to eliminate the speaker dependant characteristics by computing the cepstral coefficients. The cepstrum is computed to suppress the source signal. This can be interpreted as the spectrum of a spectrum. Speaker dependant harmonics of the fundamental frequency are transformed to one higher order cepstral coefficient under ideal conditions.

5. The typical MFCC feature vector is calculated from a window of 512 sample points and consist of 13 cepstral coefficients, 13 first and 13 second order derivatives. This reduces the dimensionality from 512 to 39 dimensions.

Different MFCC features are used for experimentation in the classification of comprehension levels in Chapter 5.

2.4.5 Laughter Detection

Laughter detection is not discussed in this thesis as this is outside the scope of the thesis due to the complexity and difficulty in detection during the analysis of features used for classification.

2.5 Video Summarisation

In this section we provide an overview of Video Summarisation. We introduce the two key approaches to video summarisation and identify the key issues with both approaches (Truong and Venkatesh, 2007).

There are two basic approaches to video abstraction or summarisation: Key frames and Video Skims. Key frames are also called *representative frames* or *still-image abstracts*, and consist of a set of salient images extracted from the underlying video source. Video Skims, also called *moving-image abstract* or *summary sequence*, consists of a collection of video segments extracted from the original video. Segments are then joined together. It is itself a video clip but of a significantly shorter duration than the original. One possible example of a video skim is a movie trailer, i.e. The *Jigsaw* movie trailer is an example of a video skim of the *Jigsaw* movie. It is important to note that a movie trailer is not always an example of a video skim, for example, if a movie trailer has additional background music added then it is not strictly a skim of the original movie.

One advantage of video skims over key frames is the ability to include audio and motion elements that can enhance the expressiveness and information in the abstract. Also, it is often much more entertaining and interesting to watch a skim rather than a slide show of images. On the other hand, key frames are useful for organising for browsing and navigation purposes.

2.5.1 Key Frame Abstraction:

Key techniques for the abstraction of key frames from video are reviewed in (Patel and Rajput, 2014) and presented below. Key frame abstraction is one of the core techniques for video abstraction. We now review and discuss whether this is suitable for the work described in this thesis.

Visual Attention Clues: A Visual Attention System model was used for key frame video summarisation by (Peng and Xiao-Lin, 2009). The Attention Detection System is divided into two parts: static attention and dynamic attention. Static attention detection is based on human attention to an interesting background even with no motion, and dynamic attention is based on human visual attention deduced from local and global motion. A Visual Index Descriptor (VID) based on a visual attention model is used for mapping between low level concepts and high level concepts. The Optical-Flow algorithm from Lucas-Kanade is used for key frame block mapping. Key frame is divided into 8*8 blocks. Final attention models are produced from combination of static attention and dynamic attention and computer weights of each attention.

Adaptive Association Rule Mining (ARM) Adaptive ARM is a technique

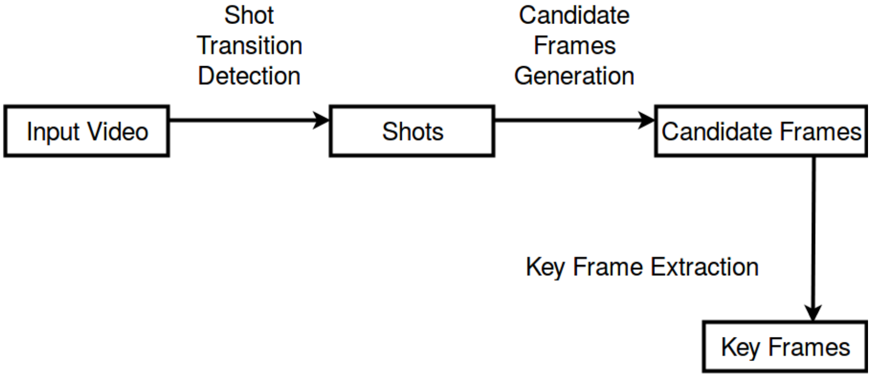


Figure 2.12: Key Frame Extraction - This is a re-drawn version of the image from (Nasreen and Shobha, 2013)

used for mining events from video (Zhang et al., 2013). ARM bridges a gap between Near-Duplicate key frames and high level semantic concepts. It has been utilised in finding the associations of the visual features of near-duplicate key frames. ARM is comprised of three steps:

1. Data Mining - In the data pre-processing step, near duplicate key frames are extracted from video and irrelevant key frames removed.

2. Adaptive Association Rule Mining - In the Adaptive Association Rule Mining step, important terms and their frequencies are calculated and combined into groups using transitive closure.

3. Classification - In the classification step, correlations between grouped terms and near-duplicate key frame groups are expressed in the form of a matrix and labelled as a class.

Motion Focussing Method: Within each video shot, this method focuses on one constant-speed motion and aligns video frames by fixing this focused motion into a static situation. The method generates a summary image containing all moving objects, and embedded with spatial and motional information, (Li et al., 2009). Four steps involve:

1. Background subtraction is applied to extract the moving foreground for each frame.

2. Using the first few frames, parameters for image alignment are estimated.

3. Initial foreground summary image is constructed. Binary segmented images are scaled and shifted with parameters gained from step 2, and mosaiced together to form the foreground summary. The min-cut method (Boykov and Kolmogorov, 2004), from graph theory, a minimum cut of a graph is a cut that is minimal in some sense, is used to find the changes to the objects in different frames. This is performed after scaling and transformation by using the least square method.

4. Find local optimal solution to problem, using as few binary segmented images as possible to cover no less than 95% foreground region in the foreground summary image.

Final output summary image provides a whole impression of all moving objects present in the video, and also the spatial and motional relationships between objects that are not captured directly by the camera.

Summarisation Based on Depth and Colour Information: This uses a colour camera combined with a depth camera to detect events. Depth information can resolve the light and shadow change problem and works in three steps (Chou et al., 2013):

1. Background Extraction: The image Depth and Colour is captured from the camera and integrated with the image. Image background is updated so that foreground and background are shown cleared.

2. Foreground Extraction: The Object is detected from the image based on depth. If depth is more, colour information is used to find the foreground.

3. Suspicious Event Detection: Object detected by one-way and two-way crossing line detection. Alarm of two-way will be triggered when crossing the warning line no matter from which side, so just a warning line is set and not the direction. Alarm of one-way not only detects whether the moving objects cross

the warning line but also detects the direction of moving objects crossing the warning line.

2.5.2 Video Skim Abstraction:

Video abstraction techniques were reviewed in (Li et al., 2001). In this section, we focus on their review of video skimming techniques. As one of the core strategies of video abstraction, we review key techniques, and then discuss the potential benefits of applying this approach to our work on presentation summarisation. Figure 2.13 demonstrates video skimming techniques.

The Informedia Project at Carnegie Mellon University (Smith and Kanade, 1995; Hauptmann and Smith, 1995; Smith and Kanade, 1998) aimed to create a very short synopsis of original video by extracting significant audio and visual information. Text keywords were extracted from manual transcripts and close captioned. Audio skimming was created by extracting the audio segments corresponding to selected keywords as well as including their neighbouring segments for improved comprehension. Following this, image skims were created by selecting video frames which were either a) frames with faces or text, b) static frames following camera motion, c) frames with camera motion and either human faces or text, or d) frames at the beginning of a video scene, with a descending priority. Following this, a video skim was generated by analysing the word relevance and the structure of the prioritised audio and image skimming. Experiments demonstrated impressive results on some types of documentary video with explicit speech or text contents. However, such results may not be achievable on videos containing more complex audio contents.

Siemens Corporate Research (Toklu et al., 2000) reported work on video skimming where multiple cues were employed, including visual, audio and textual information. Detected shots were first grouped into story units based on detected areas of 'change of speaker' or 'change of subject'. Audio segments which correspond to generated story units were then extracted and aligned with the summarised closed-caption texts. Representative images were also extracted for each story unit from a set of key frames consisting of all first frames from underlying shots. A shot is an uninterrupted sequence of frames between cuts. The final video skim includes audio and text information, yet the key frame information is excluded. This approach depends heavily on textual information, and involves obtaining text summaries at different level of detail, selecting key frames that best represent the visual track, and then using shot grouping results to eliminate all key frames belonging to an anchorperson or reporter shot.

(Nam and Tewfik, 1999) from University of Minnesota generated skims based on a dynamic sampling scheme. A continuous video source was first decomposed into a sequence of 'sub-shots' - intra-shot boundaries introduced into individual shots. The motion intensity index - a procedure for calculating the amount of visual activity within each 'sub-shot' unit, is then computed for each of them. To obtain a localised sampling rate, the time axis was partitioned into discrete regions of different sampling rates. The motion intensity index curve was quantised and converted to different sampling rates with a pre-defined sampling table. The full-rate sequence was sampled in each time interval at the corresponding sampling rate determined in previous step. Finally, a piece-wise constant frame interpolation scheme was used to render the true temporal nature of the video. Thus, given a summary sequence and corresponding information on local sampling rates, intermediate frames were produced between successive retained frames by copying the preceding frame sample as many times as desired.

(Hanjalic and Zhang, 1999) of Microsoft Research clustered all video frames into n clusters, with n varying between *1* and *N*. They then performed cluster-validity analysis to determine the optimal number of clusters for the given data set. The next step was to build a set of key frames. For this, one representative frame was chosen from each of the clusters and taken as a key frame of the sequence. Chosen for this purpose was the cluster elements closest to cluster centroids. This formed the final key frame sequence. Each shot, to which at least one extracted key frame belongs, is taken as a key video segment. These key segments are then concatenated to form the preview. While theoretically this method can be applied to a video sequence of an arbitrary length, the sequences of interest in this work are constrained to specific events with well defined and reasonably structured content. The reason for this constraint is that long video sequences are mostly characterised by an enormous visual content variety, which is difficult to classify in a number of distinct clusters and, consequently, difficult to represent by a limited number of key frames/segments. Therefore, in order to be able to efficiently apply this approach to a full-length movie, it is necessary to first segment the movie into well-structured, high-level fragments. In video, a shot is a series of frames that run for an uninterrupted period of time.

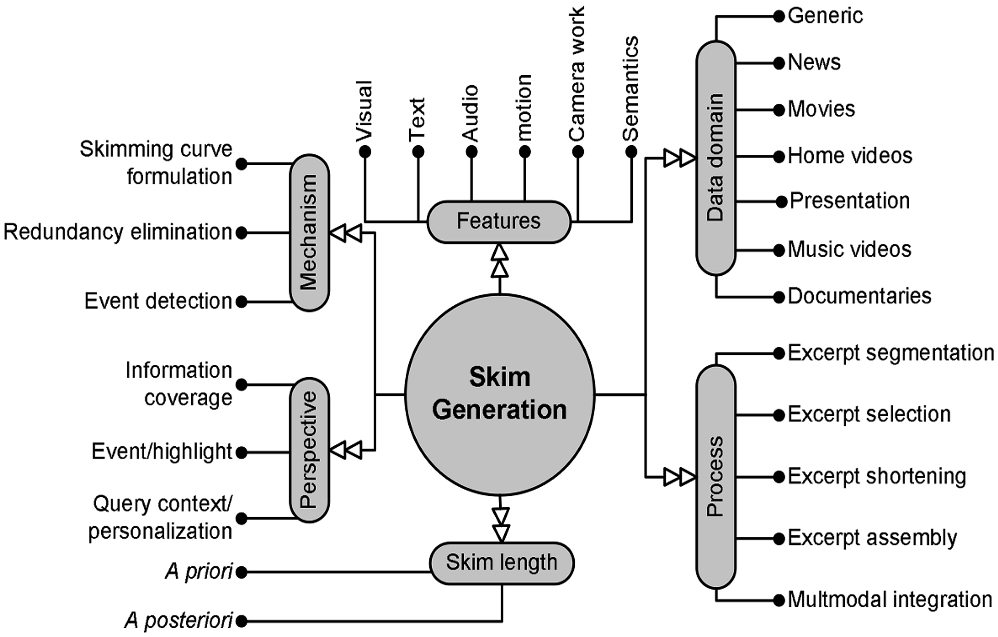


Figure 2.13: Attributes of Video Skimming Techniques (Truong and Venkatesh, 2007)

Further work from Intel Corporation, (Lienhart, 1999), focused on the summarisation of home videos considered more usage model-based than content-based. The date and time of the recordings were first extracted. Following this, all shots were clustered into five different levels based on the date and time taken. The five clusters include the individual shots, the sequence of contiguous actions where the temporal distance between shots are within five minutes, the sequence of contiguous actions where the temporal distance between shots are within one hour, the individual days and the individual multi-day events. Next, a shot shortening process was performed where longer shots are segmented into 2-minute long video clips. To select desired clips, the audio signal was calculated and employed in the selection process based on an observation that during important events, the sound was more clearly audible over long periods of time than it is with less important content.

2.5.3 Discussion

For the work described in this thesis, where we wish to summarise academic presentations in which much of the information exists in the audio stream, key frames are of little use due to the missing important audio information from these summary types. We focus on generating video skims of these presentations which can capture all important parts of the presentations and summarise to the most interesting content for the user. In the following section we describe the general video skim generation process.

2.5.4 Skim Generation Process

The following is a generic process for automatically generating video skims: excerpt segmentation, excerpt selection, excerpt shortening, multimodal integration and excerpt assembly (Truong and Venkatesh, 2007). The term excerpt refers to a segment of video, be it a shot, scene or event spanning a number of shots etc. In practice, a skim generation technique may skip certain steps or combine them in different variations. However, the principle and essence remains as described for all video skimming works.

Excerpt Segmentation This step segments the whole video sequence into separate units. However, the segmentation of the video sequence into anonymous segments is considered *a priori* process rather than part of the skim generation. The segmentation of speech can be done by detecting pauses in speech by using the time-code of the extracted transcript. Excerpt selection is essentially the process of dividing the video sequence into event and non-event segments.

Excerpt Selection The next step is to select the excerpts to be included in the skim. Coverage of generated skims will depend on the technique used for excerpt selection and will influence the context and coherence level of the skim. Sometimes the skim length may need to be taken into account when selecting events. Excerpt selection can also be done recursively by first forming scenes and clusters of shots.

Excerpt Shortening The shortening procedure ensures that the excerpt is concise without loss of information. Shortening an excerpt runs the risk of creating inappropriate cut points, reducing overall coherence and irritating the viewer. The simplest method for shortening an excerpt is to select a predetermined portion of the excerpt. However, this step is generally omitted if the summarisation perspective is highlighting, allowing the user full comprehension of highlighted parts of the video sequence.

Multimodal Integration and Excerpt Assembly The generated excerpts discussed so far are often single-modal, normally audio, image or textual. The purpose of this step is to insert the missing modality, realign excerpt boundaries and combine all excerpts in to the final skim. If done correctly, multimodal integration can enhance the coverage, context and coherence of the produced skim. Video skims can be classified into two types based on the integration of audio and video information: modal synchronisation and modal asynchronisation. In the former, the audio and video streams are synchronised according to the timeline of the original video sequence.

2.5.5 Related Works at TRECVID

The development of new approaches for accessing relevant content in multimedia data is a core challenge at TRECVID run by the National Institute for Standards and Technology (NIST) (Smeaton et al., 2006). The TREC conference series is sponsored by NIST with the goal of encouraging research in information retrieval. In 2001 and 2002, TREC sponsored a video track to encourage research in automatic segmentation, indexing and content based retrieval of video. This track became an independent evaluation in 2003. TRECVID has included a wide variety of tasks including ad-hoc Video Search, Instance Search, Multimedia Event Detection, Surveillance Event Detection, Video hyperlinking, Concept localisation, and Video to text description.

- A new ad-hoc video search task was introduced to model the end-user video search use-case, who is looking for segments of video containing persons, objects, activities, locations and combinations of such. The ad-hoc video search task is as follows: Given a standard set of shot boundaries for the test collection, and a list of 30 queries, return for each query the top 1000 video clips from the standard set, ranked according to the highest possibility of containing the target query, (Awad et al., 2016).
- An important need in many video collections is to find video segments of a certain specific person, object, or place. The task for the systems is as follows: Given a collection of test videos, a master shot reference, a set of known location / scene example videos, and a collection of topics, locate for each topic up to 100 clips most likely to contain a recognisable instance of the person / object / place in one of the known locations, (Over et al., 2014).
- A user searching for events and complex activities occurring at a specific place and time involving people or objects in multimedia material may be interested in a wide variety of potential events. A technology is needed that can take as input a human-centric definition of an event that developers and systems can use to build a search query, (Awad et al., 2016).
- The video hyperlinking task requires the automatic generation of hyperlinks between given manually defined anchors within source videos and target videos. The results of the task for each anchor is a ranked list of target videos in decreasing likelihood of being about the content of the given anchor, (Awad et al., 2016).

- The localisation task challenges systems to make their concept detection more precise in time and space. In this task, systems are asked to determine the presence of the concept temporally within the shot. For each concept from the list of ten designated for localisation, a list of up to 1000 clips where each video shot may or may not contain the concept. For each I-Frame within each shot in the list, systems were asked to return the x,y co-ordinates of the upper left and lower right vertices of a bounding rectangle, (Awad et al., 2016).

Typical approaches applied at TRECVID include the early and late fusion, see Section 2.6.3, of a number of low-level audio and visual feature descriptors. The most common low-level features applied include a combination of some of the following: Scale Invariant Feature Transform (SIFT), Section 2.3, ColourSIFT, Colour Moment Grid (CMG) - measures that characterise colour distribution in an image, Edge Orientation Histogram (EOH) - edge orientation is evaluated by searching the maximum response over a set of edge filter kernels, SIFT with Hessian-Affine descriptor, SIFT and hue histogram with dense sampling, Histogram of Gradients (HOG) with dense sampling - HOG actual gradient direction is calculated and then binned, Dense trajectory with Motion Boundary Histogram (MBH) feature Improved Trajectories Feature (IDT), OpponentSIFT, RGB-SIFT, RGB-SURF - see Section 2.3, ORB descriptor - see Section 2.3, and Motion Boundary Histograms on both x and y directions (Peng et al., 2014; Inoue et al., 2014; Douze et al., 2014).

In addition to the above low-level features, some mid-level features are also commonly applied including some or all of the following: HMDB51 attributes - a dataset of 7,000 video clips of 51 basic action classes (dive, jump etc.). ImageNet attributes - dataset of 1.2 million images, each prominently representing one object from a total of 1,000 object classes (Douze et al., 2014). Convolutional Neural Network (CNN) - Another 4,096 dimensional features extracted using the same convolutional network trained on the ImageNet 2010 data (Douze et al., 2014). Some approaches also adopt high level feature descriptors including Optical Character Recognition (OCR) - see Section 2.3.3, and LIMSI Automatic Speech Recognition

(LIMSI ASR) (Douze et al., 2014).

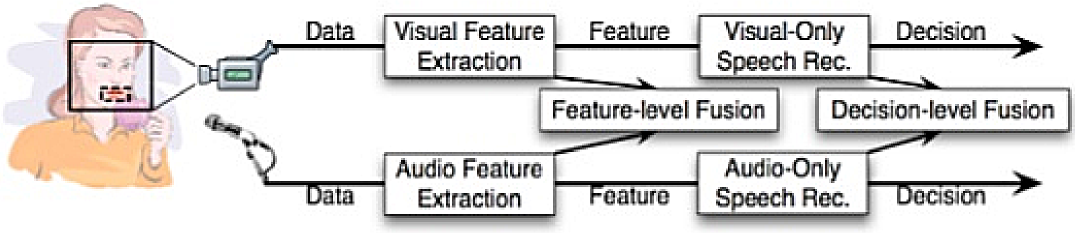
Gaussian Mixture Models (GMMs) are often used to model video shots. GMM's are often used for data clustering, GMM's cluster by assigning query data points to the multivariate normal components that maximise the component posterior probability. The GMM parameters are estimated for each shot under the maximum a posteriori (MAP) criterion. GMM Supervectors are then extracted by combining normalised mean vectors. Typically, linear Support Vector Machines (SVMs) are trained with a combination of some of the above described feature extractors. SVMs are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Some systems apply an early fusion approach while others apply a late fusion approach (Peng et al., 2014; Inoue et al., 2014; Douze et al., 2014).

The works above utilise MFCCs, see Section 2.4.4, which are the only audio based features utilised. MFCCs have limitations such as the assumption that the fundamental frequency is lower than the frequency components of the linguistic methods. MFCCs also have a lack of interpretation, meaning their reaction to accented speech or noise is unknown. One of the most common visual features used are dense trajectories - in which trajectories are obtained by tracking densely samples points using optical flow fields, which, although performing well in Human Action Recognition tasks are very time consuming. The works described above use SVMs for classification. SVMs have some well-known limitations such as its limitation in choice of kernel, the problems presented by discrete data, and slow performance in test phase (Lutter, 2015; Hao et al., 2013; Burges, 1998).

2.6 Multimodal Signal Fusion

The nature of the research conducted in the thesis requires the fusion of different multimodal signals for the training of classifiers to predict high level concepts such as audience engagement, speaker ratings, and audience comprehension. The section looks at some of the common issues and methods associated with multimodal signal fusion. The definition of multimodality as used in this thesis is for signals of different modalities originating form the same device, such as speech signal, presenter visual signal, presentation slides etc.

Multimodal Signal Fusion is concerned with finding the optimal solutions to the problems associated with the fusion of signals from each of the various modalities. Some of the common multimodal signals which can be fused include the following: keypoint detection, SIFT features / ORB detector, colour histogram, optical flow for motion, gesture heat maps, face and head detectors, edge and corner detectors, pitch, intensity, speech rate, articulation rate, MFCC's etc. The most common problems associated with this fusion of signals from each of the separate modalities are the questions of when to fuse the separate modality signals, how to fuse, which level of fusion is best suited for a particular domain and what features to fuse. This section reviews and outlines some of the open issues in multimodal signal fusion as reviewed by (Atrey et al., 2010).



Data Level:
e.g., combining 2 webcam video streams, multiple perspectives

Feature level:
e.g., combining speech and lip movements

Decision level:
e.g., combining gestures and speech

Figure 2.14: The 3 basic levels of fusion - from (Dumas et al., 2009)

2.6.1 Fusion Methods

There are different fusion methods which have been used by many researchers to perform multimodal analysis. These methods fall into the following three categories which are summarised below. Figure 2.14 from (Dumas et al., 2009) shows the three basic levels of fusion.

Rule-based Methods These include a variety of basic rules of combining multimodal information including statistical rule-based methods such as linear-weighted fusion, MAX, MIN, AND, OR and majority voting. There are also custom-defined rules constructed for the specific application perspective. Rule-based schemes generally perform well when there is good temporal alignment between modalities.

Classification-based Fusion Methods These are used to classify multimodal observation into pre-defined classes. These include Support Vector Machine, Bayesian Inference, Dempster-Shafer theory, Dynamic Bayesian Networks, Neural networks and Maximum Entropy Model. Of these methods, Bayesian Inference and Dynamic Bayesian Networks are generative models while SVMs and neural networks are discriminative models.

Estimation-based Fusion Methods These include the Kalman Filter, extended Kalman filter and particle filter fusion methods (Atrey et al., 2010). These have primarily been used to better estimate the state of a moving object. For example during object tracking multiple modalities are fused to better estimate the position of the object.

2.6.2 When to Fuse

In this section we describe when to fuse data from different modalities such as audio and video. When the fusion should take place is an important consideration in the multimodal fusion process. Data from different modalities is usually captured in different formats and at different rates and therefore need to be synchronised before fusion. As data fusion can be performed at the feature as well as at the decision level, the issue of synchronisation is considered at these two levels. Due to different time periods of data processing and feature extraction, the question of when the features should be combined remains an issue. Simple strategies can be to fuse the features at regular intervals, although this strategy may not be the most appropriate it is computationally less expensive. Alternative strategies may be to combine all features at the time instant in which they are all available.

2.6.3 Level of Fusion

One of the first considerations in multimodal signal fusion is the question of which strategy to take when fusing the information. One of the most widely used approaches is to fuse the information at the feature level which is known as early fusion. Another approach is known as decision level or late fusion which fuses the separate modalities in the semantic space. A hybrid approach can also be adopted which is a combination of these approaches.

- **Early-Fusion** In early fusion the features extracted from input data are combined and then sent as input to the analysis unit that performs the analysis.

- **Late-Fusion** In late fusion the analysis unit first provides the local decisions that are obtained based on individual features. Local decisions are then combined using a decision fusion unit to make a final decision.

- **Hybrid Multimodal Fusion** To exploit advantages of both early and late fusion, several researchers have opted for a hybrid fusion strategy, a combination of both early and late fusion techniques.

2.6.4 What to Fuse

The separate modalities used in a fusion process may provide complementary or contradictory information, therefore understanding which modalities are providing complementary information is required. This is also related to finding the optimal number of media streams or feature sets to accomplish an analysis task. If the most suitable streams are unavailable, can we use alternative streams without loss of cost-effectiveness or confidence?

Figure 2.15 demonstrates different fusion levels, (Atrey et al., 2010). Specifically, feature level or early fusion are shown in b and d. Decision level or late fusion is shown in c and e. A Hybrid approach to multimodal analysis, taking advantage of both feature level and decision level fusion is demonstrated in f.

2.7 Summary

In this chapter we looked at previous work related to the topics investigated in this thesis. We reviewed work in the area of non-verbal communication and looked at audio and visual feature extraction. We conclude that there is limited use for visual feature extraction tools due to the domain of the presentations, in which the main information exists in the audio stream. We make use of some basic visual extraction tools such as Facial Detection, Optical flow and Optical Character Recognition for recognising movement in the presenter and the audience, and for detecting information form the presentation slides. We primarily make use of audio feature extraction tools for this work given that the main information exists in the audio stream. These include Pitch and Intensity extraction, Speech formants F1 - F4, and MFCC features.

We then reviewed the main approaches to video summarisation and multimodal signal fusion. We conclude the most appropriate summarisation approach for work of this type is a video skim, which are effectively highlights of the original video. Key frame extraction is of little use in this work due to the importance of the audio track. We also discuss the different fusion levels of data level fusion, feature level fusion and decision level fusion. The next chapter introduces the dataset collected and annotated for work described in this thesis.

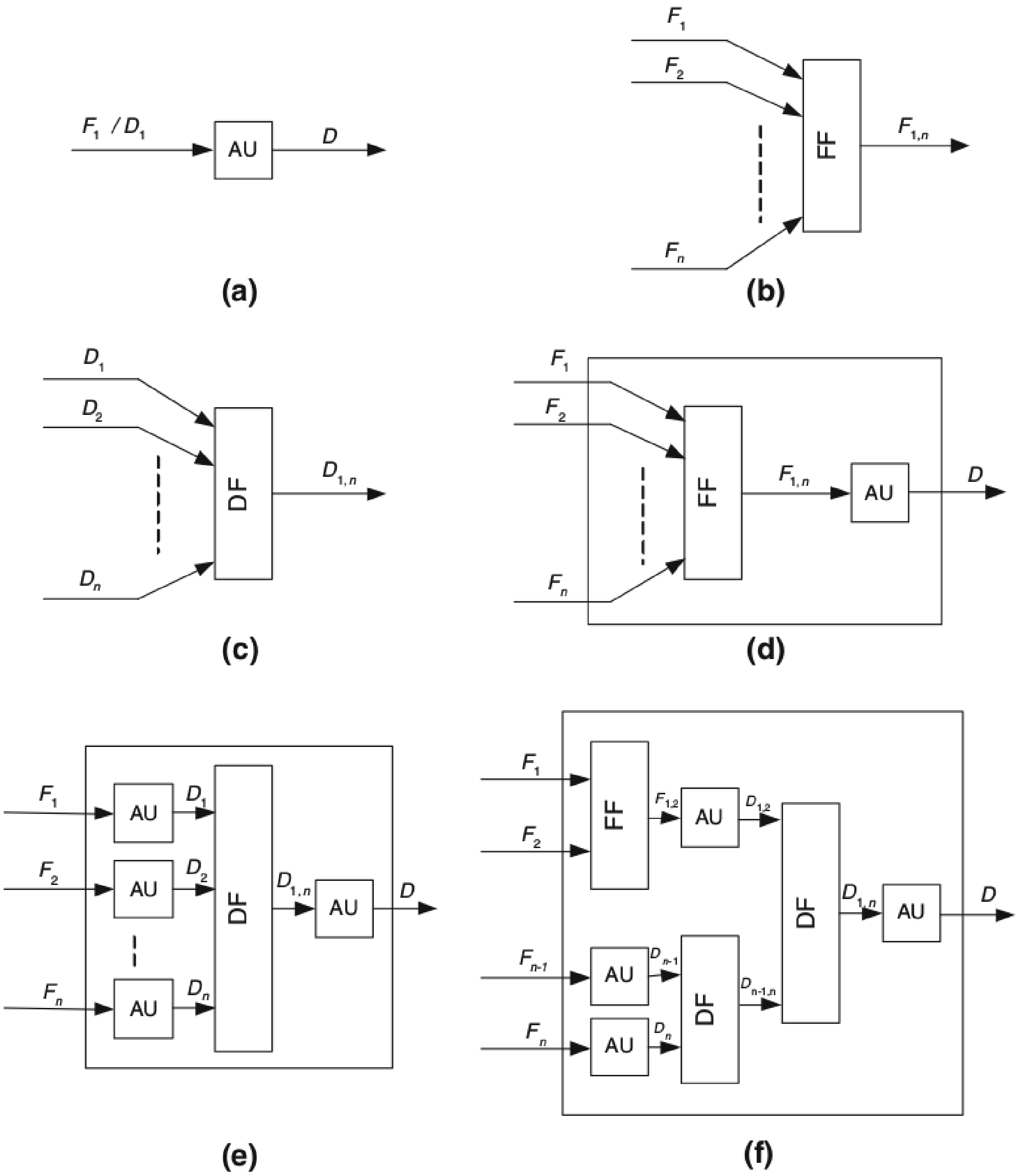


Figure 2.15: Multimodal Fusion: Strategies and Conventions. a) Analysis Unit b) Feature Fusion Unit c) Decision Fusion Unit d) Feature Level Multimodal Analysis e) Decision Level Multimodal Analysis f) Hybrid Multimodal Analysis (Atrey et al., 2010)

Chapter 3

Multimodal Data Collection

There are a number of existing multimodal datasets in the public domain, however, none of these are suited to the research undertaken in this thesis. In order to obtain a suitable dataset for the work described here, we recorded a multimodal dataset of academic presentations from a conference held in Dublin. We collected recordings of the speaker and of the audience in addition to presentation slides. Annotations were also carried out on this newly collected data. The data, the data collection process, and full annotation schemes are described in detail in this chapter. We begin by overviewing some of the existing publicly available datasets, indicating why they are not suitable for this investigation.

We describe the collection of the dataset used in the investigations reported in this thesis. We describe the collection methods and overview the details of the collected dataset. We then describe the manual annotation of the dataset with features to enable the investigations necessary to address the research questions introduced in chapter one.

3.1 Existing Multimodal Datasets

This section gives a brief introduction to several existing multimodal datasets.

3.1.1 AMI

The AMI meeting corpus, (Carletta et al., 2005), is a multimodal dataset consisting of 100 hours of meeting recordings. It is based on a set of pseudo design project meetings, where the team sit around a table, with a supporting white board or interactive screen, and describe their design ideas. One third of the corpus is completely natural recordings of meetings which would have taken place whether recorded or not. The remaining two-thirds of the corpus is comprised of participants playing different roles in fictitious meetings. The AMI corpus includes high quality manually produced transcriptions for each speaker. It also contains a range of other annotations including dialogue acts, topic segmentation, extractive and abstractive summaries, named entities, head gesture, hand gesture, gaze direction, movement around the room and emotional state etc. It contains wide camera views, overhead camera views and individual close-up camera views.

As the research described in this thesis in based on conference presentations rather than meetings, with typically just one or two presenters speaking in front of a large audience, we consider this dataset unsuitable for addressing the research questions introduced in Chapter 1 of this thesis.

3.1.2 MASC

The Manually Annotated Sub-Corpus (MASC), (Ide et al., 2010), consists of approximately 500,000 words of contemporary American English written and spoken data drawn from the Open American National Corpus[1]. MASC includes manually validated annotations for sentence boundaries, tokens, lemma and part of speech; noun and verb chunks; named entities and discourse structure. It also has a 100K+ sentence corpus with WordNet 3.1 sense tags. MASC contains a balanced selection of texts from a broad range of genres. It is a large scale, open, community based effort to create needed language resources for NLP. Audio-visual material is not

[1] http://www.anc.org/

included in this dataset.

As this corpus does not contain audio-visual material it is not suitable for our work described in Chapter 1 of this thesis.

3.1.3 Open Online Lectures and Presentations

Many lectures and presentations are freely available online from sources such as *TED Talks*[2] and *Coursera*[3] platforms and also those available on *YouTube*[4]. The difficulty with using these date sources for our research is that no corresponding video exists for the audience to these lectures and presentations.

3.2 Introduction to Corpus

Since no existing available dataset is suitable to enable us to address the research questions introduced in Chapter 1, it was necessary for us to develop our own dataset. This dataset must consist of parallel recordings of both the presenter and the audience to academic presentations. We need a natural corpus with a number of academic presentations, all with an audience present and recorded. The Speech Prosody 7 conference[5], an international conference related to speech, held in Trinity College Dublin in May 2014 provided an ideal opportunity for us to collect a suitable dataset.

After gaining ethical clearance from the host university, all presenters at the conference were asked to give permission for the recording of their presentation(s). Also, all attendees to the conference were asked to give their approval for the recording of the audience to the academic presentations.

This conference had several different presentation types, including keynote talks, research presentations from different sections, and poster presentations. For this

[2] https://www.ted.com/talks

[3] https://www.coursera.org/

[4] https://www.youtube.com/

[5] http://fastnet.netsoc.ie/sp7/

dataset, we required at least 30 uninterrupted recordings of full academic presentations, which must be recorded in good quality, and have fully synchronised recordings of the audience to each presentation. Recordings needed to have a full view of the stage, in addition to recordings which focus on just the slides being presented, to act as back up for PDF versions of each presenter's slides.

3.3 Data Collection

We recorded presentations at *Speech Prosody 7* using three SONY HDR-XR500 cameras. Video was recorded in 1080p at 29.97 fps with H264 codec. Audio was recorded in Dolby Digital 48kHz, 16 bit stereo at 256 kbps. The recording standard used was AVCHD. PDFs and back-up recordings from each presenters slides were also recorded.

A total of three fixed cameras were used to record the corpus. Two cameras were fixed within the gallery at the approximate mid-point of the seating structure. One camera being set to record the overall wide-angle view of the whole stage, including the presenter, slides and the surrounding stage area. The other camera zoomed in to record the presentation slides in order to provide a back-up to those provided to us by the presenters.

A final, third camera was set up just behind and slightly to the side of the presenter in order to record the audience during each presentation. This gave full recordings of the presenter and of the audience to each presentation. Presentation recordings were later synchronised by matching presentation recordings and audience recordings using the acoustic footprint - points of high volume in both videos (speaker coughing, clapping etc.) were aligned by analysing the audio signal. This allowed us to properly align presentation and audience videos for annotation purposes. High quality audio was recorded separately and later synchronised to the video recordings. Recordings were processed by (Spoken Data Video Processing), by running fully trained Automatic Speech Recognition (ASR) over the presentations

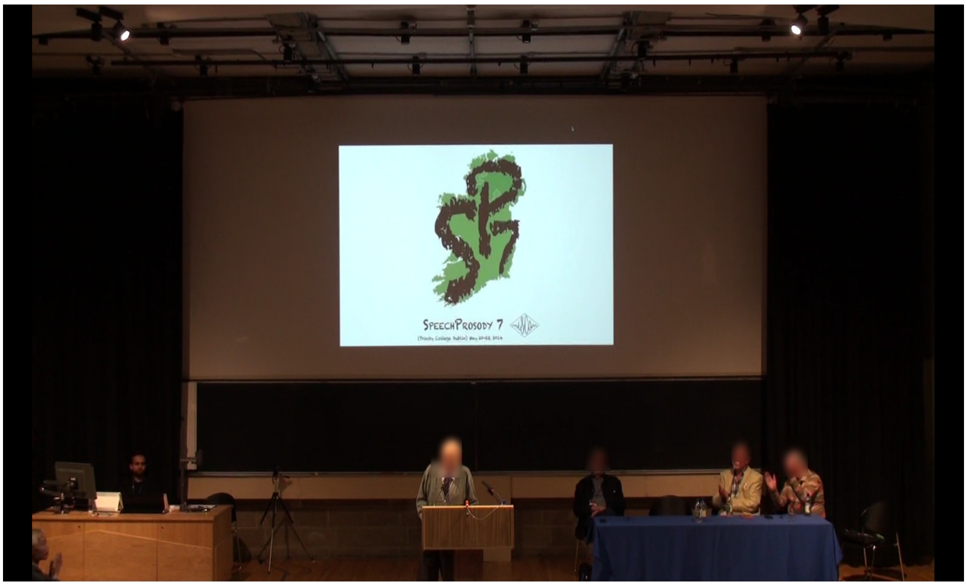


Figure 3.1: Full camera view of the stage

and outputting full speech transcripts and keywords for each presentation recording. Data used in this research is available online at (Super Lectures Video Hosting)[6].

Some presentations were deemed unsuitable for human annotation and experimentation due to a number of factors - The four keynote talks were felt to be too long, and were observed to have too little change in audience engagement levels over the course of the presentation. These presentations were excluded from further processing. The final corpus consists of dual video recordings from the presenter and audience of 31 scientific talks, totalling 520 minutes of conference video. This gave us a total of 1040 minutes of video requiring human annotation for each paralinguistic concept studied.

The final videos were used from (Super Lectures Video Hosting) for presenter based annotation. Full presentation videos from *Spoken Data* and slide recordings were passed to *Super Lectures*, videos were then processed to include aligned slide changes and were produced to production quality H.264 codec in mp4 format. These videos had a frame rate of 25 fps and a bit-rate of 768 kbps. All had been processed with MPEG-4 AAC audio codec with a sample rate of 44100 Hz and an audio bit-rate of 86 kbps.

[6] `https://www.superlectures.com/` is a video hosting website dedicated to conference videos and lectures. Videos can be watched on tablets, smartphone, laptop or desktop

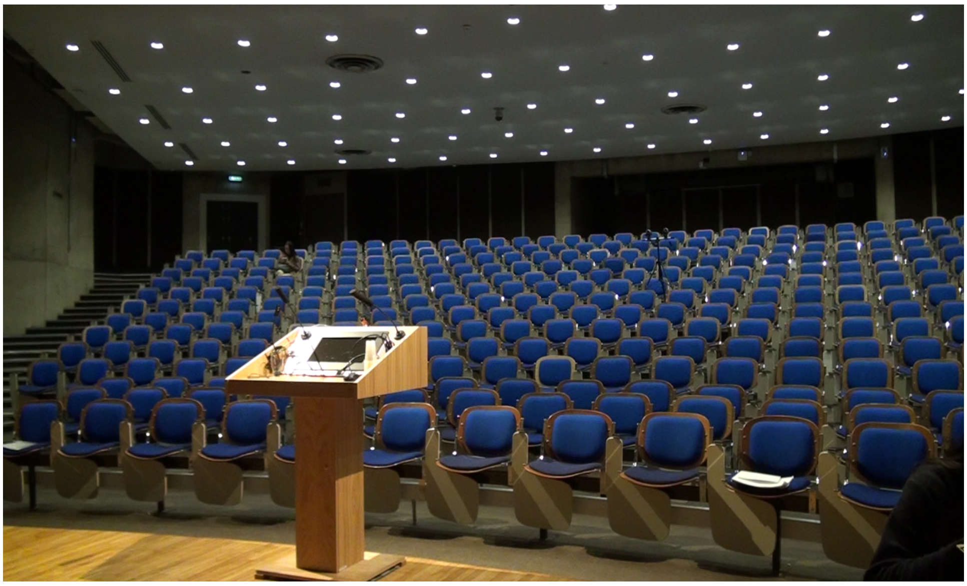


Figure 3.2: Full camera view of the gallery

3.4 Human Annotation

In order to make this dataset useful for our research, human annotations were required to label the paralinguistic features of emphasis, speaker ratings & engagement and comprehension. To facilitate this, we developed an annotation tool to perform assessment of the levels of engagement and comprehension, or in the case of spoken emphasis, to judge whether or not the content is emphasised. These manual labels were used as the gold-standard for the paralinguistic features of the presentations studied in this thesis.

Human annotators of speaker ratings, audience engagement and emphasis tasks were recruited from a pool of research students, support staff and research engineers across Dublin City University and Trinity College Dublin. These consisted of an equal balance of native English speakers and non-native speakers. Some, but not all annotators, had prior experience of work on spoken content.

Presentation videos were uploaded to *YouTube*, from which video segments were embedded into the annotation tool. This tool played video segments, selected by choosing the set of segments annotated the least number of times, and then choosing a random segment from this set. Annotation records from each annotator were recorded which later allowed us to analyse annotator ratings for consistency.

3.4.1 Speaker Quality & Audience Engagement

Our objective was to obtain gold-standard labels for speaker ratings and audience engagement levels. In order to determine suitable lengths of content for annotation we performed a pre-study with a small number of subjects. In this, the subjects were asked to watch a selection of video segments, ranging from 10 seconds up to 50 seconds. Participants were asked to select the best segment length based on time taken to make judgements of engagement levels within the audience, whilst avoiding segments that were too long and thus allowing too much change to occur in engagement levels. If too much change occurs in engagement levels during an annotation segment, estimating the level of engagement will be more inconsistent and less meaningful. The 30-second video segments were selected as the best based on the results of the pre-study. Audience and presenter video segments were of the same length and times.

To create labels for speaker quality and audience engagement, annotators were asked to watch 30-second video segments, selected at random from the collection, and to estimate the audience engagement level for this video segment based on an ordinal scale from 1 to 4. Prior to performing the engagement rating, participants were provided with example labelled video segments from each of the 4 engagement levels. Annotators were also requested to provide an estimate of the attendance level at each talk, estimated on a scale from 1 to 5. For a full auditorium, this would be rated as 5, where as for an empty auditorium this would be rated as 1.

Following this annotation task, participants were asked to watch 30-second video segments of the speaker, different segments than they viewed the audience for, selected at random, and to rate the speaker according to their level of agreement with the following statement ‘This is a good speaker who is able to capture the attention of the audience and bring the presentation to life.’ Annotators were asked to base their judgements on both acoustic and visual stimuli. Human judgements provided

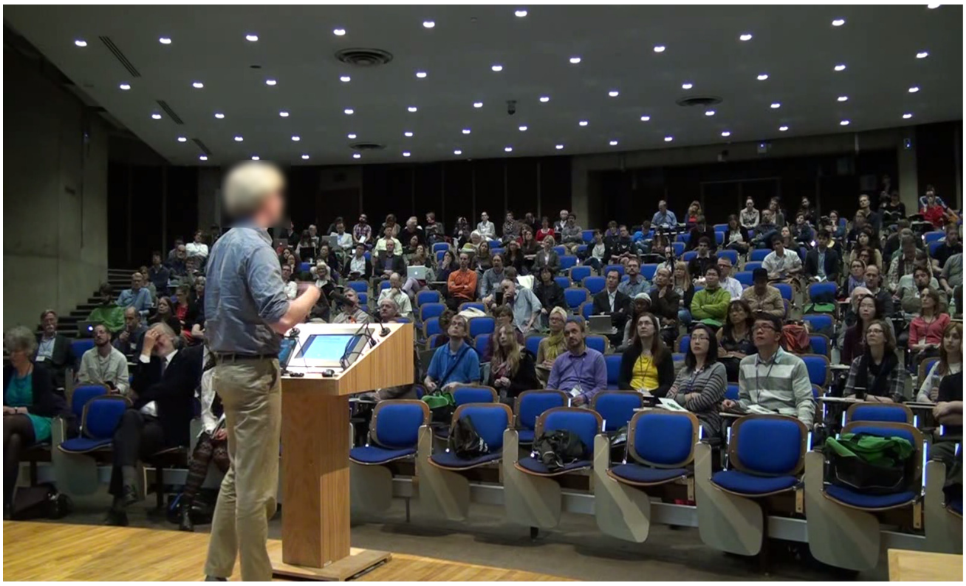
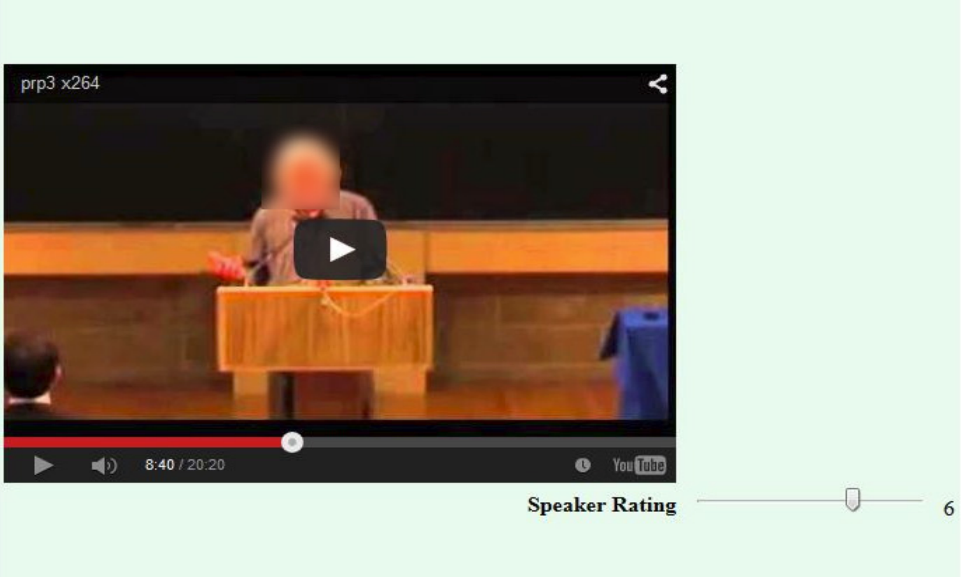


Figure 3.3: Audience View - View of the audience to presentations, this is the view annotators had for judging audience engagement levels

were made on a Likert scale from 1 to 8, with 1 being the weakest level of agreement with the given statement and 8 being the strongest level of agreement. Views of the presentation slides were excluded from the annotators view in order to ensure that human judgements were based solely on the strengths of the speaker and not on the content.

Segments were allocated to annotators at random. First, the set of segments with no annotations was selected. Next, a segment was selected at random from this set. Audience and presenter video segments were of the same length and times. The 30-second segment length was selected from a pre-study performed within the research group. In this, a number of researchers were asked to watch a selection of video segments, ranging from 10 seconds up to 50 seconds, and to select the best segment. Participants were asked to select the best segment length based on time taken to make judgements of engagement levels within the audience, whilst avoiding segments that were too long and thus allowing too much change to occur in engagement levels. If too much change occurs in engagement level during an annotation segment, estimating the level of engagement becomes a much more difficult task for the annotators, which in turn affects the reliability of the annotations.

Figure 3.4: Presenter Close-up view - View of the presenter up close, this is the view that annotators had for annotating emphasised speech.



Speaker Rating 6

Figure 3.5: Image of the Annotation Tool used for human annotation of speaker ratings.

For annotation we follow the assumption, as observed from watching the dataset, that audience engagement levels will not vary too much over a short period of time. As each segment was annotated only once, the labels may not be consistent due to annotator bias, for example, individual annotators may consistently rate engagement higher or lower than each other. In order to reduce this potential bias, a number of steps were taken to prepare the data by smoothing the human annotations.

- **Outlier Removal:** The first step involved the removal of obvious outliers from the dataset which were re-annotated by different annotators. These outliers were defined as labels with significantly different values to nearby segment annotations. An example of an outlier of this type occurred if we had a sequence of video segments receiving an engagement rating of 4, followed by a segment receiving an engagement rating of 1, these outlier segments were re-annotated immediately by different annotators from the pool.

- **Normalisation:** The next step involved the normalisation of labels. This was achieved by analysing ratings for each annotator and applying either a lowering or highering effect to annotator labels to bring each annotator's ratings in line with other annotations. For example some annotators were found to have an annotation range from 2 to 5 while others were found to have a range from 3 to 7. By analysing annotations we were able to match these up and lower annotations which were on the high side or increase annotation ratings which were found to be on the low side.

- **Time Windowing:** The next step involved Time Windowing, this was performed in order to apply smoothing to annotations and reduce the effect of annotator bias. Video segments were aligned into time windows each 90 seconds in length and in steps of 30 seconds. In order to find the label for each 90-second time window, we took the mean of labels for each video segment within that time window. This resulted in annotations for three sequential video segments being combined and averaged.

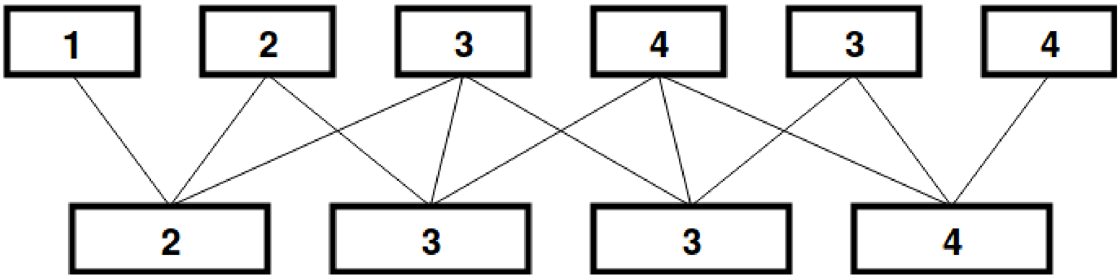


Figure 3.6: Example of Time Windowing of Video Segments

3.4.2 Speaker Emphasis

The next task was to obtain human annotations for intentional or unintentional emphasised speech in academic presentations. Ten human annotators were divided into two groups of five. Annotation of speaker emphasis in the audio-visual presentations was carried out in two phases. We asked the first group of annotators to watch two five minute clips from separate audio-visual presentations and to mark areas of the video where they perceive the presenter to be applying emphasis either intentionally or unintentionally. In order to obtain gold-standard annotations for audience engagement at a fine-grained level, these annotators were also asked to watch two 5-minute clips from the audience to different presentations and to estimate audience engagement levels for 6-second video clips. This was to investigate any correlations between emphasised speech and audience engagement.

While engagement annotations in the previous section were over 30-second segments, the purpose of engagement annotations in this section is to investigate potential correlations with emphasised speech, hence we chose a shorter 6-second segment length so that potential correlations with emphasised speech may be investigated.

The second group of annotators watched the presentations and labelled emphasis for the video clips corresponding to those for which the first group labelled audience engagement at a fine-grained level. This group also labelled audience engagement levels for the video clips which had been labelled for emphasis by the first group. This ensures that a total of four separate 5-minute video clips, each from different presentations and given by different presenters, were labelled both for emphasis and

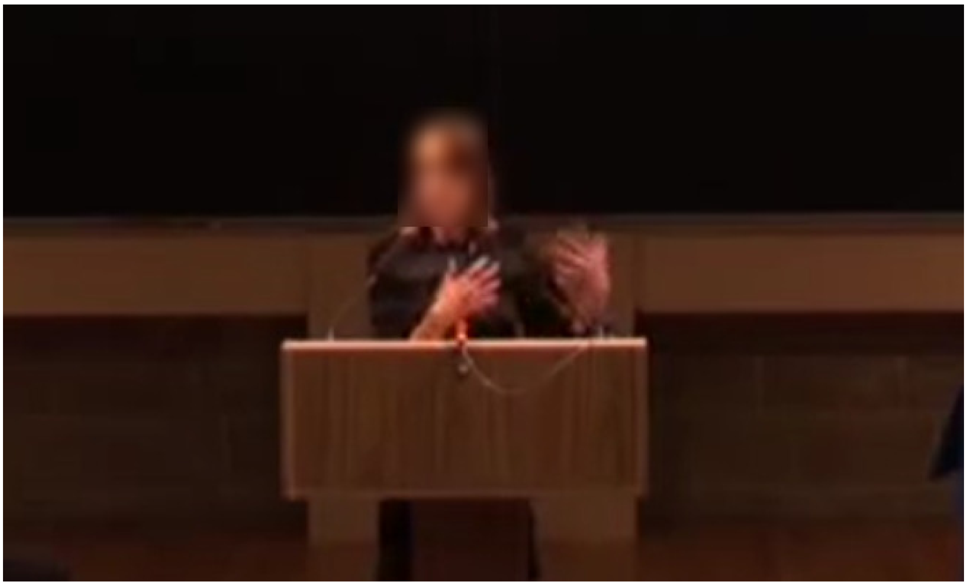


Figure 3.7: Presenter: mid-Emphasis

fine-grained audience engagement, and that these concepts were labelled by different annotators to avoid annotator bias. A total of 10 annotators were recruited for this emphasis annotation task, and were each paid 5 euro after completion of tasks. As a result, each 5 minute clip was annotated for each concept by 5 different annotators.

Engagement levels were annotated in this study on an ordinal scale from 1 to 6. This is a finer scale than was used for classification of audience engagement, which was annotated on a scale from 1 to 4. This is to allow for investigation of correlations between emphasised speech and audience engagement. Most engagement annotations from the previous section were found to be in the range between 1 and 3. Thus, it was felt we needed a wider range of labels to properly investigate potential correlations between emphasis and audience engagement. The average of each segment's 5 separately judged engagement annotations is taken as the gold-standard label for audience engagement.

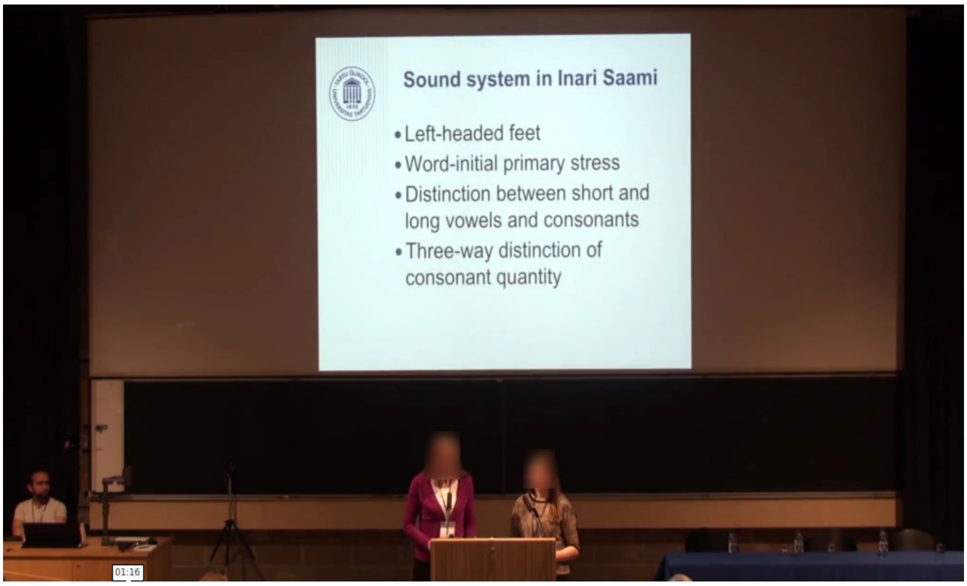

Sound system in Inari Saami
Left-headed feet
Word-initial primary stress
Distinction between short and long vowels and consonants
Three-way distinction of consonant quantity
01:16

Figure 3.8: Two presenters jointly present a talk

3.4.3 Comprehension

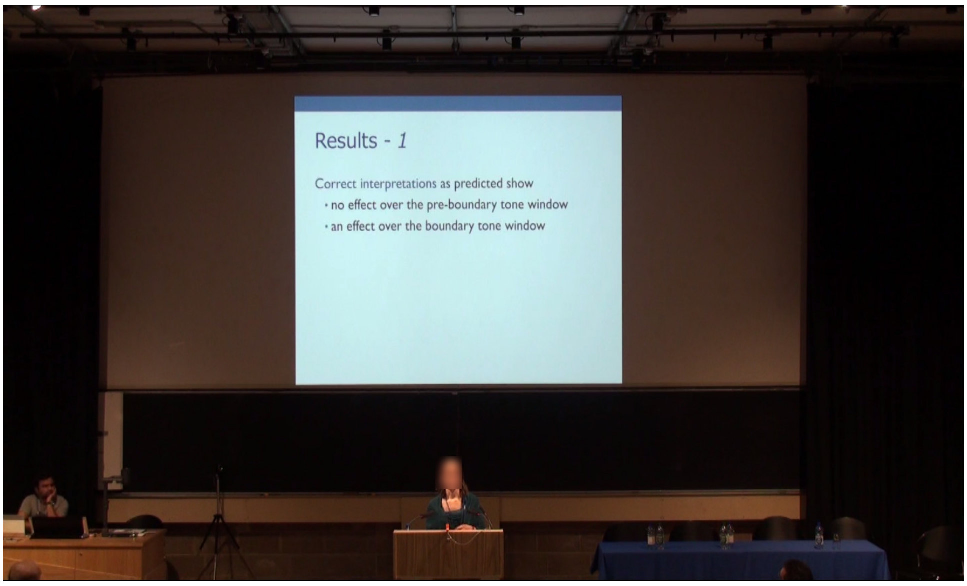
In order to study the concept of audience comprehension we first needed to obtain gold-standard labels for comprehension levels over the dataset. For this task each of the 31 academic presentations in the dataset was divided into between 4 and 7 contiguous video segments. Segmentation decisions were based on changes of topic within presentations. Each video segment was between 2 and 4 minutes in length. This gave a total of 172 video segments requiring human annotation. Human annotation of comprehension levels was required over each video segment in our dataset. Annotators were required to watch a full presentation, by watching each segment of a presentation in order of occurrence. The first task required them to provide a text summary of the current segment, following this they were required to provide an estimate of just how much they comprehended the content of that segment.

The purpose of written summaries was to have our annotators think about the content first before providing their comprehension estimate and also to provide a means to ensure quality of annotations. Following this, annotators were asked to provide an estimate of how comprehensible they considered the material to be on an ordinal scale from 1 to 8. An even numbered scale was chosen in order to encourage our annotators to make a decision on comprehension level rather than choosing the middle, neutral option.

We consider this to be the best way to annotate for a concept as abstract as comprehension, since to estimate comprehension levels of those attending each talk in person would require a questionnaire completed by each person in attendance at the presentation. This was not realistic for the corpus used in this study.

Annotation of comprehension was performed by human annotators recruited from a popular crowdsourcing website. Annotators were paid an average rate of 7.50 euro per hour. Recruited annotators all had English as their first language and all had at least some third-level education. Annotators each watched contiguous audio-visual segments from one full academic presentation. Each video segment was annotated by at least three annotators and the final gold label was calculated from the average of the three annotations. A total of 93 paid annotators were recruited to perform annotations and the quality of their work was checked before payment issued by studying their provided text summaries and comparing with their estimated comprehension levels. For example, if an annotator was unable to provide an accurate text summary of the presentation they had just watched, then it was unlikely they could have a high level of comprehension for that segment.

To assess the level of inter-annotator agreement for this task we calculated the intra-class correlation model 1, ICC(1,1), over all annotations, which assumes that annotators rating different subjects are different, being subsets of a larger set of annotators, and chosen at random (Shrout and Fleiss, 1979). The intra-class correlation was calculated using the online ICC calculator available at (Chinese University of Hong Kong). A set of guidelines was published on selecting and reporting intra-class correlations (Koo and Li, 2016). The authors state that typically values of less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability. The mean ICC(1,1) reliability score for this work on comprehension was found to be 0.6034, which taking into



Results - *1*

Correct interpretations as predicted show

- no effect over the pre-boundary tone window
- an effect over the boundary tone window

Figure 3.9: Annotators View: This is the view which annotators had for judging presentation rankings and for performing comprehension annotation tasks.

consideration the subjectivity of the task at hand, we regard this to be a good level of agreement between the assessors.

3.5 Final Dataset

The final annotated dataset consists of parallel recordings of presenter and audience for 31 academic presentations, fully annotated for speaker ratings, audience engagement, audience comprehension levels and partly annotated for emphasis. A PDF version of presentation slides and back-up recordings of these exist for each presentation. Three of the presentations were given jointly by two speakers, while two presenters each gave two presentations. This gave a total of 32 individual presenters, 13 of these were male and 21 of which were female. Native English speakers accounted for 11 of the 32 individual presenters. 9 of the 31 presentations used were given in full by a native English speaker, while 1 of the 3 presentations jointly given by 2 presenters consisted of 1 native and 1 non-native English speaker. The other 2 of these jointly presented talks were given by two non-native speakers. All other presentations were given in full by non-native speakers.

The 31 fully annotated presentations included in this dataset have been subdivided into 5 categories: 5 Intonation (Inton) videos, 9 Plenary Oral (Plen) videos, 6 Speech Rhythm and Timing (speechRT) videos, 6 Perception and Production (PrP) , and 5 Slavic Prosody (slavicP) videos. This information is summarised in Table 3.1. All presentations are available for viewing on the *Super Lectures*[7] website. Figure 3.10 provides a screenshot from this website. As can be seen from the screenshot, when viewing a presentation, its abstract is displayed above the video, with information on slide changes and the current slide also available.

3.6 Summary

This chapter describes the collection of a suitable dataset to investigate the research questions described in chapter 1 of this thesis. Human annotation of the paralinguistic features of speaker rankings, audience engagement, intentional or unintentional emphasised speech and audience comprehension is also described. In the following chapters of this thesis, we perform automatic classification of these concepts in presentation videos and use these classified concepts to automatically generate video summaries of the presentations in this dataset.

We have attempted to get human assessors as representative of the original audience as possible for annotation tasks and final summary evaluation tasks. Human annotations for speaker ratings, audience engagement, and speaker emphasis were performed by research students and research staff at Dublin City University and Trinity College Dublin. Around half of these human annotators had experience in working with spoken content. Human annotations for audience comprehension were crowd-sourced, with restrictions in place specifying that annotators needed at least some level of third-level education, in order to get crowd-workers as close as possible to the original audience. Finally, human participants for eye-tracking evaluations of presentations summaries, described later in Chapter 6, were recruited from research students at Dublin City University, many of whom were students of

[7] `http://SuperLectures.com/`

the School of Applied Linguistics, making this group of assessors very close to the original audience who witnessed the presentations used in this work.

Table 3.1: Data Collection Summary

	Inton	Plen	PrP	slavicP	speech	Total
presentations	5	9	6	5	6	31
By 2 Speakers	1	1	0	0	1	3
male presenters	3	3	3	2	2	13
female presenters	3	7	3	3	5	21
native speakers	2	7	1	1	0	11
non-native speakers	4	3	5	4	7	23

Table 3.2: Final Annotated Engagement Levels

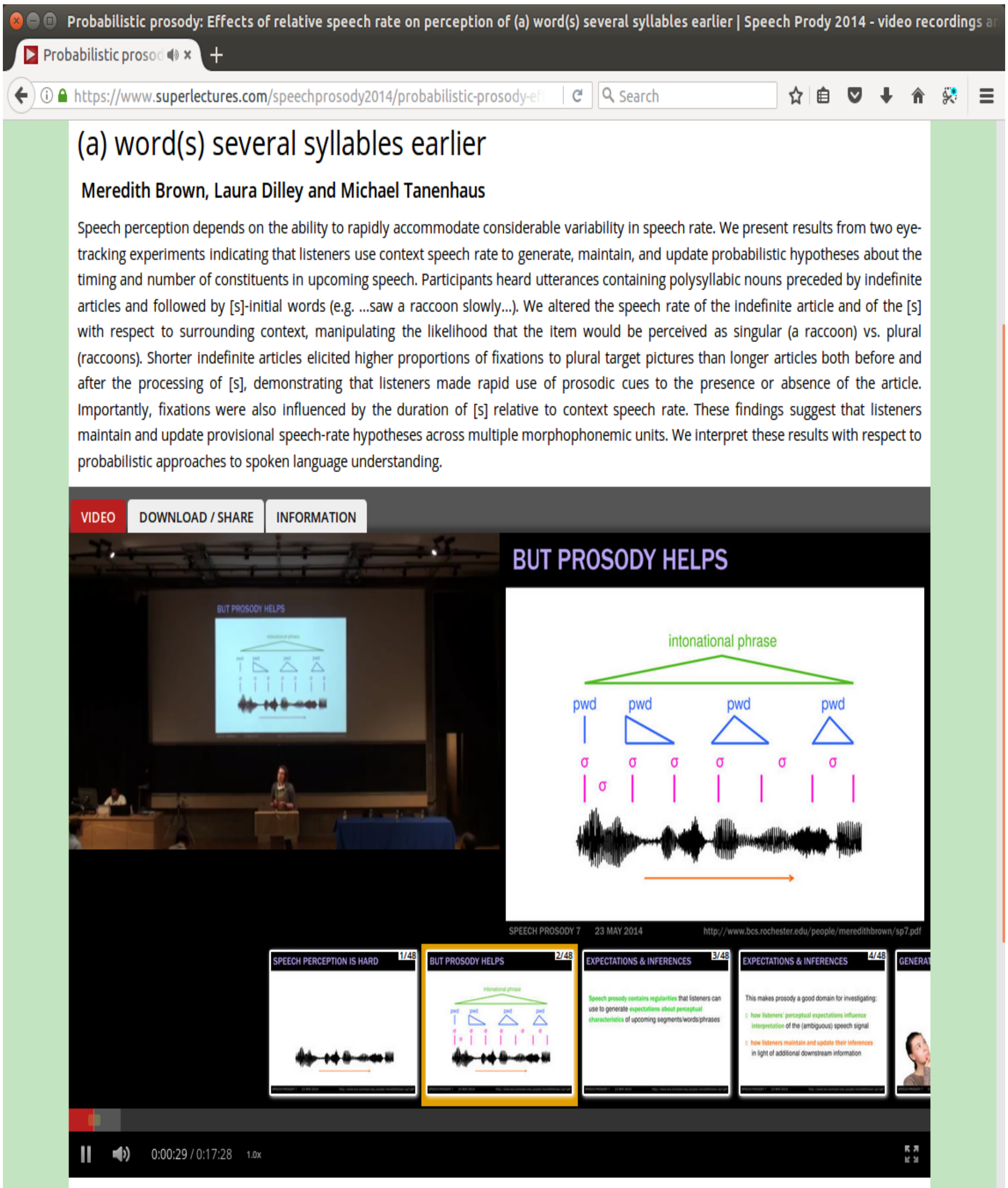
Engagement Level	1	2	3	4
Number of Segments	1	190	610	176

Table 3.3: Speaker Movements : Audience Engagement - Linear Correlations

Label	Label	$r =$
Speaker Movements	Audience Engagement	**0.187705**

Table 3.4: Final Annotated Comprehension Levels

Comprehension Level	1	2	3	4	5	6	7
Number of Segments	2	40	50	36	23	11	10



Probabilistic prosody: Effects of relative speech rate on perception of (a) word(s) several syllables earlier | Speech Prody 2014 - video recordings a

https://www.superlectures.com/speechprosody2014/probabilistic-prosody-ef

(a) word(s) several syllables earlier

Meredith Brown, Laura Dilley and Michael Tanenhaus

Speech perception depends on the ability to rapidly accommodate considerable variability in speech rate. We present results from two eye-tracking experiments indicating that listeners use context speech rate to generate, maintain, and update probabilistic hypotheses about the timing and number of constituents in upcoming speech. Participants heard utterances containing polysyllabic nouns preceded by indefinite articles and followed by [s]-initial words (e.g. ...saw a raccoon slowly...). We altered the speech rate of the indefinite article and of the [s] with respect to surrounding context, manipulating the likelihood that the item would be perceived as singular (a raccoon) vs. plural (raccoons). Shorter indefinite articles elicited higher proportions of fixations to plural target pictures than longer articles both before and after the processing of [s], demonstrating that listeners made rapid use of prosodic cues to the presence or absence of the article. Importantly, fixations were also influenced by the duration of [s] relative to context speech rate. These findings suggest that listeners maintain and update provisional speech-rate hypotheses across multiple morphophonemic units. We interpret these results with respect to probabilistic approaches to spoken language understanding.

SPEECH PROSODY 7 23 MAY 2014 http://www.bcs.rochester.edu/people/meredithbrown/sp7.pdf

Speech prosody contains regularities that listeners can use to generate expectations about perceptual characteristics of upcoming segments/words/phrases

This makes prosody a good domain for investigating:
- how listeners' perceptual expectations influence interpretation of the (ambiguous) speech signal
- how listeners maintain and update their inferences in light of additional downstream information

Figure 3.10: Screenshot from Super Lectures website which hosts all presentations used in this research.

Chapter 4

Classification of Engaging & Emphasised Material

This chapter investigates the classification of speaker ratings and audience engagement within academic presentations. We also investigate the classification of intentional or unintentional emphasised speech in such presentations. The purpose of this chapter is to provide the first exploratory work into the goals of this thesis which is to provide a novel method for summarising academic presentations using high-level paralinguistic features. This chapter addresses three research questions, the first two of which are introduced in Section 4.1.1, the third research question is introduced in Section 4.2.1.

4.1 Classification of Automatic Speaker Ratings and Audience Engagement

In the first part of this chapter we report on our investigations into the classification of speaker ratings for presenters of academic presentations. We investigate audio-visual features best associated with high presentation ranking and with high audience engagement, we then investigate the use of these speaker ratings to aid in the building of a classifier for classification of audience engagement.

4.1.1 Introduction

We all have ideas about what constitutes a good public speaker. While some of these can be purely subjective, others are generally universal, such as maintaining eye-contact with the audience, avoiding monotone speech, and avoiding speaking too quietly. A speaker who lacks energy and is unable to connect with their audience is unlikely to capture the interest of their audience, even if their presentation topic is otherwise interesting. On the other hand, we speculate that an energetic presenter who gestures frequently, making eye contact with different members of the audience, and alternates their speaking tone depending on context, timing etc., is likely to capture much greater interest in their presentation.

Switching to the presentation itself, we have all been present at interesting presentations. As well as the presenter capturing the attention of the audience through their use of techniques mentioned above, the content must hold some interest. The presenter can seem to induce a certain energy in the room, this itself can at times seem to be contagious, with the audience reacting to the presenter through nodding, laughing etc.

Most of us have also been present for presentations where the speaker lacks energy and is unable to bring the presentation to life. This results in presentations where the audience quickly loses interest and the content fails to capture their interest. Such presentations can have the effect of putting the audience to sleep, and even if the topic should generally be of interest for the audience, a poor presenter can be difficult for the audience to follow.

In this chapter we investigate the potential to automatically rate the abilities of each presenter, that is, how a human would rate each presenter's effectiveness at giving presentations. Details of the human annotations and the dataset used for these investigation were introduced in the previous chapter. In this chapter we study speaker quality with audience engagement for presentations given at an academic conference. We explore the features of speech which humans consider to be good speaking techniques and investigate the development of an automatic classifier to label these automatically.

We study the relationship between good speaking techniques and levels of audience engagement for scientific talks. In order to define ‘good’ speaking techniques for this study, a pool of annotators rated the strengths of speakers for our experimental dataset. We analyse correlations between the identified features of good speaking techniques and levels of audience engagement. We explore the development of a classifier to automatically rate the strengths of a presenter within an audio-visual recording of a presentation. We then investigate the extent to which these ratings can be used to predict the level of engagement of the audience with the talk in progress.

In the first part of this chapter we address the following research questions introduced in Chapter 1:

1. **‘Can we build a classifier to automatically rate the qualities of a ‘good’ public speaker?’**

2. **‘Can we build a classifier to automatically predict the levels of audience engagement by utilising speaker-based and basic visual audience-based modalities?’**

We hypothesise that by extracting a set of influential audio-visual features from video of the presenter we will be able to classify speaker ratings for each presenter, similar to those assigned by human annotators. We also hypothesise that the speakers use of speaking techniques shown to influence audience engagement in conjunction with analysing reactions of the audience, can be used to predict audience engagement levels during presentations of this type.

We investigate the potential correlations between individual audio-visual features extracted from the dataset and a presenter’s effectiveness at giving presentations and audience engagement levels. We also investigate potential correlations between speaker ratings themselves and end audience engagement levels.

4.1.2 Objectives

In order to be able to automatically classify the quality of an audio-visual presentation, we first need to extract features from the recorded data which can be used as the input to such a classifier. In order to do this, in the first part of this chapter we examine the features which we can extract from the data which are likely to be reliable indicators of presentation quality. Firstly considering audio features, we hypothesise that pitch and intensity of the audio signal are likely to be highly influential. The range of these features is also likely to be important as well as individual and average values, in addition to information regarding the variance of values. Thus, we calculate the high, low, mean and standard deviation for each of these features.

With respect to visual features, gesticulation on the part of the presenter is highly likely to be of importance in the classification of speaker quality. Without the use of specialised equipment during the recording of a presentation, such as the *Microsoft Kinect*, which can track gestures, tracking and identification of individual gestures is a difficult task and one that is outside the scope of this thesis. Instead, we make use of Optical Flow to track the total motion around the area of the presenter. This has the effect of tracking total gesturing by the speaker, as well as head movement and changes in body posture.

Additionally, we use the Face Detector available through *OpenCV*[1] to track the presenter's head movement. Once again, the range of these features is likely to be important as well as the absolute values. Thus, we calculate the high, low, mean and standard deviation for optical flow and facial detection features.

In addition to these we also use Optical Flow to estimate the total movement of the audience during presentations. We hypothesise that during regions of presentations which the audience do not find very engaging, there is likely to be quite a lot of movement within the audience as people can have a tendency to either get restless (much movement), or alternatively get sleepy (no movement), during pre-

[1] https://opencv.org/

sentations in which they are not engaged. In addition to this, we hypothesise that during interesting presentations, audience members are more likely to be focussed on the presentation, attending to the presenter or looking at displayed slides, than during unengaging presentations, where they are likely to lose focus and look around a lot. To this end, we also look to extract an average facial count from the audience for front facing participants as an indication of the audience's level of focus on the presentation.

We also wish to find a way to estimate the attendance at each presentation, as this is important when interpreting the extracted facial counts and motion from the audience. For example, in a presentation with a small number of attendees, we would expect a reduced degree of motion within the audience and a smaller number of front facial counts than for a presentation with a much larger number of attendees. Where the presentation with the lower attendance shows similar figures to the one with a higher attendance for facial counts, this could indicate that this is actually an interesting presentation from the audience's perspective.

The following section goes into more detail regarding the extraction of these audio-visual features.

4.1.3 Multimodal Feature Extraction

A set of acoustic and visual features was extracted from the multimodal corpus of scientific talks. Acoustic and visual features from the presenter were used for our speaker rating classifier and audience engagement classifier, while visual features of the audience were extracted for our audience engagement classifier. Acoustic features were extracted using Praat (Boersma et al., 2002), explained in more detail in Appendix C.1.3, while visual features were extracted using OpenCV by (Bradski and Kaehler, 2008), explained in more detail in Appendix C.1.1.

4.1.3.1 Speaker-Based Acoustic Features

We now describe the audio features extracted from the speaker which were used to train a classifier.

- **Pitch** values were extracted using AutoBi Pitch Extractor (Rosenberg, 2010). We use default min and max values of 50 and 400 respectively as pitch values rarely went outside of this range. We used the max, range, mean and standard deviation of extracted pitch values for our experiments.

- **Intensity** was extracted using AutoBi Intensity Extractor (Rosenberg, 2010). Similar to pitch, we used the max, range, mean and standard deviation of Intensity values. This generated an Intensity contour using default parameters of a minimum Intensity of 75dB and a time-step of 100ms. This minimum intensity level was chosen to exclude overly noisy data.

- **Speech Rate**, defined as the number of syllables per duration - the total utterance time from to start to finish, and **Articulation Rate**, defined as the number of syllables per phonation time - total time spent phonating words, were extracted for each presentation using the Praat script (De Jong and Wempe, 2009). This script extracts the number of syllables, number of pauses, duration and phonation time from an audio file. These values were used to calculate the speech rate and articulation rate.

4.1.3.2 Speaker-Based Visual Features

We now describe the visual features extracted from the speaker for the training of a classifier.

- **Head movement** was extracted using Robust Facial Detection (Viola and Jones, 2004). For this task we used a publicly available pre-trained cascade for head & shoulder detection (Castrillón-Santana et al., 2008), to detect the presenter's head and return the (x,y) coordinates for the top-left corner of a bounding rectangle for the location of the speaker's head at that point in time. We then calculated total head movement per frame by taking the Euclidean distance between (x,y) points in corresponding frames. We calculated the mean and standard deviation of head movement per second to return a standardised measure of speaker head movement. For this task we used the head & shoulder cascade rather than the standard frontal face cascade, also made publicly available by the same authors, this detects faces which are facing to directly frontwards, as the standard frontal face cascade is likely to fail if the speaker turns to the side. As with the audio features, the use of the mean value gave us the presenter's average head movement per second, while standard deviation accounted for variation in this. We use this feature to investigate the effects of the speaker's head movement during a presentation on audience engagement.

- **Speaker Motion** was extracted using an optical flow coded implementation in OpenCV (Lucas et al., 1981). We calculated the total pixel motion changes from frame to frame to put more weight on directional changes in motion than on continuous motion to account for changes of direction, and take the mean and standard deviation in overall speaker motion. We did this as we considered movement from one side to another then back again as being more significant than movement just going from one side to another. The mean of this value gave us an average of the total motion per second, while standard deviation accounts for variation of motion. This was achieved by returning the total value of calculated motion from frame to frame. All frame change values in a one-second interval were added and the mean calculated to get the speakers average motion per second, and the corresponding standard deviation calculated for variation in these values. This value then accounted for all movement on the part of the speaker, including head movement, body movement and gesturing. The isolated measure for head movement as described above was used to normalise this figure to estimate for gesticulation alone.

4.1.3.3 Audience-Based Visual Features

We now describe the visual features extracted from the audience for the training of a classifier.

- **Face Counts** were extracted using a facial detector (Viola and Jones, 2004) on a trained frontal face cascade, provided by (Castrillón-Santana et al., 2008), over the audience video to detect faces that were facing forwards towards the speaker or the slides. This used the same configuration as for Head Movement of the speaker, except for the pre-trained cascade used. For this we used the frontal face cascade which is likely to fail as the person turns their head to the side. This is because all we wanted for this feature was an estimate of the number of attendees facing directly forwards, towards the presenter and their slides. This method was based on the hypothesis that positive frontal face detections were more likely if the person is facing forwards. The total number of positive detections per second was counted, and the mean and standard deviation of these values calculated per video segment. This gave an indication of the percentage of the audience observed to be paying close attention to the presentation. Interpretation of this depends on the overall attendance figure as explained below.

- **Audience Motion** was extracted in a similar way to Speaker Motion described above. As above we took the mean and standard deviation of the audience level motion. This gave an estimate of total motion within the audience, and again depended on the overall attendance figure as explained below. This estimate of total movement within the audience was used to estimate audience engagement levels.

- **Audience Attendance level** - We developed an algorithm for estimation of attendance at talks, as follows.

  1. Attendance was estimated by first taking a vacant seat as a region of

interest and calculating High and Low values of Red, Green and Blue from the region of interest. This method assumes all seats are of the same colour. In the multimodal corpus used for this study all seats were blue.

2. Pixels which had values falling within this range were counted which could then be used to give an estimate of attendance.

3. The next stage involved recording the high and low values of the number of pixels falling within this range over the entire corpus, this was then taken as the low and high level of attendance. This is explained by the assumption that the highest number of vacant seats equals the lowest attendance level, and the lowest number of vacant seats equals the highest attendance level. The mean of all samples from each talk was then calculated as the final figure for that talk.

4. As our lowest attendance figure will not be equal to zero attendance, as no talk is attended by exactly zero people, and our highest attendance figure will not equal full attendance, as no talk has 100% attendance, which would be assumed when provided with an attendance estimate of 1 out of 5, or 5 out of 5 respectively, we next need to calculate a normalisation value to make final outputs more accurate. For this, the mean attendance value calculated from each talk was divided by the range between high and low values to give an added_value figure. An integer value of 4, representative of the range from low and high attendance estimates 1 to 5, was then divided by this added_value figure, and the result incremented by 1. This was then subtracted from our highest attendance value of 5. Finally for our normalisation value, we divide this added_value figure by 5, representative of the attendance range.

$$NormalisationValue = (5 - ((4/AddedVal) + 1)/5)$$

5. Add this result to the initial added_value figure. Finally, in order to calculate attendance estimates as a percentage figure, this normalisation value was divided by the average attendance figure multiplied by 100.

   $$AttendEstimate = ((NormalisationValue+AddedVal)/(AvgAttend*100))$$

6. To evaluate this method for estimating attendance, we took the average attendance provided by our human annotators as our ground truth and evaluated the closeness of our estimates with this ground truth. All human annotations on attendance level were averaged to give the groundtruth attendance level between 1 and 5 as labelled by our human annotators. By selecting a region of interest only once at the start of the video collection we achieved an overall accuracy figure of 88.04%.

Algorithm 1 Estimate Attendance

```
$AttendanceEstimate \rightarrow AE$
$_1 SelectvacantseatasR.O.I.$
$_2 CalculateHighandLowRBGvaluesfromR.O.I.$
for all $Presentations \rightarrow P$ do
    for all $_3 MinutesofPresentation \rightarrow M$ do
        $AE \leftarrow PixelswithinRBGHighandLow$
        $Total \leftarrow Total + 1$
    end for
    $AE \leftarrow AE/Total$
    $_4 CalculateNormalisationValueNV$
    $_5 AE \leftarrow AE + NV$
end for
$_6 FinalMethodEvaluation$
```

4.1.4 Experiment Design

In this section we explain the design of the experimental investigations performed in this chapter. The intention of this study is to investigate the automatic classification of a presenter's speaking ratings and audience engagement during presentations.

While previous work in the study of speaking techniques and prediction of engagement, as referred to in our review of related work in Section 2.2.1 and Section 2.2.2, has tended to use a regression model for classification, human annotations on speaker ratings and audience engagement in this study are rated on Likert and Ordinal scales. As pointed out by (Blaikie, 2003) and (Jamieson et al., 2004), Likert scales fall within the ordinal level of measurement. While ratings on a Likert scale have a rank order, intervals between values cannot be presumed equal. For this reason we treat the problem of classification of Likert scale ratings as an ordinal classification problem rather than as a standard regression problem, and as such, we use an Ordinal Class Classifier for classification experiments in this work, expanded upon in Section 4.1.4.1.

Although Likert scales are normally ranked over an odd numbered range, most typically 5, 7 or 9, we used an even number of classes in order to encourage our annotators to make a decision rather than just labelling segments to the middle annotation range.

Values for extracted audio-visual features are calculated over 30-second intervals, matching up with human annotated segments for speaker ratings and audience engagement as described in our Data Annotation Section 3.4.1. We chose 30-second segments for annotation of speaker ratings and audience engagement based on the pre-study we performed to assess the best segment length for annotation and classification tasks of this nature, as earlier described in Section 3.4.1.

In order to investigate the effects of individual speaker acoustic modalities on audience engagement, we calculated the Pearson Linear Correlation between extracted audio-visual features and corresponding speaker ratings and audience engagement levels. Calculations were performed over the entire dataset to discover the true correlation with human annotations.

4.1.4.1 Experimental Investigation

In this section we describe in detail the experimental investigations performed for the classification of a presenter's speaker ratings and audience engagement levels during these presentations.

We trained a classifier to automatically rate the qualities of the speaker, and estimate the audience engagement levels using Weka data mining workbench (Frank et al., 2010). Further information on Weka can be found in Appendix C.1.4. Classification was performed using an Ordinal Class Classifier (Frank and Hall, 2001) and evaluated using 10-fold cross validation. Further evaluation was also performed using Leave-One-Out cross validation. For 10-fold cross validation, the original sample is randomly partitioned into 10 equal sized sub-samples. A single sub-sample was retained as the validation data for testing the model, for which the remaining 9 sub-samples are used to train. This process was repeated 10 times with each the sub-samples used exactly once as the validation data and the 10 results from the folds averaged. Whereas for leave-one-out cross validation, a single instance was retained as the validation data to test the model which is trained on all of the other sub-samples. This process was repeated n times, where n was equal to the number of sub-samples in the data.

For classification of a presenter's speaker ratings, we used the base 8-point Likert scale as used for annotation. We then combined classes by joining consecutive classes to form a 4-point scale to investigate the effects on accuracy rates. For example, classes 1 and 2 were combined, and classes 3 and 4, etc. Audience engagement classification used the base 4-point Likert scale used for annotation. This was later reduced to a 3-class range as only one instance of the lowest audience engagement level existed after data preparation, hence the two lowest engagement classes were combined. In addition to extracted audio-visual features, final classification results for speaker ratings were used as a predictive feature for classification of audience engagement.

Linear correlations between speaker ratings and corresponding audience engagement levels were calculated to investigate the relationship between what people perceived to be good speaking techniques and the actual engagement of the audience at that time.

4.1.4.2 Detailed Technical Description

Individual speaker modalities including Pitch (max, range, mean, standard deviation), Intensity (max, range, mean, standard deviation), Head Movement (total, range, mean, standard deviation) and overall Speaker Motion (range, mean, standard deviation) were pre-fused for classification. All features were combined in a single dataset and a classifier trained on these. For audience engagement level classification, we pre-fused visual modalities of the audience including Audience Motion (mean, standard deviation) Frontal Face Counts (mean per second, standard deviation per second) and Attendance (per talk) in addition to speaker ratings and individual speaker acoustic modalities above. Individual modalities were then removed from the training set to investigate their effects on classification accuracy. This is described as pre, or early fusion, where a single classification model is trained on all used features. Audio features were extracted using *Praat*, and visual features extracted using *OpenCV*, as earlier described in more detail in Section 4.1.3.

Values for individual audio-visual features were extracted and aligned with human annotated values for speaker ratings and audience engagement levels. This was achieved by calculating averaged values over each of the 90-second time-windowed partitions of the dataset as used for human annotation of the dataset. The Pearson's Linear Correlation (Hall, 2015) was calculated from these in order to calculate the most effective individual modalities for prediction of speaker ratings and audience engagement levels. This was achieved by aligning values for each of these modalities with corresponding human annotated values for engagement and speaker ratings, for each time point within the dataset.

Following this, extracted audio-visual features were fused prior to classification in an early-fusion, or feature-level fusion approach, by combining all extracted features within a single dataset. Classification was performed over the base 8-point ordinal scale as used for annotation. These classes were then combined to form a 4-point scale for classification. Following classification of the presenter's speaker ratings, outputs from these classification experiments were then combined and pre-fused with extracted audio-visual features from the presenter and the audience for prediction of audience engagement. As described in the above section, all classification tasks were performed using an Ordinal Class Classifier, and evaluated using both 10-fold cross validation, and Leave-One-Out cross validation.

Finally, human annotations for speaker ratings and for audience engagement levels were taken and aligned over each of the 90-second time-windowed partitioned of the dataset as used for annotation, and the Pearson's correlation coefficient between these calculated. This was to discover the relation between speaker ratings and audience engagement levels, and whether one's speaking skills lead directly to high audience engagement, or if the content itself played a large or small role in resulting audience engagement levels.

Table 4.1 shows the details of the classifiers trained for classification of speaker ratings and the features used to train them. Labels for classifier type used in these and subsequent tables are defined as follows: Audio-only classifier - Classifier built only using features extracted from the audio stream. Visual-only classifier - Classifier built only using features extracted from the visual stream. Audio-visual classifier - Classifier built using features extracted from both the audio stream and the visual stream. Audience visual-only Classifier - Classifier built only using visual features extracted form the audience video. Speaker Based Features-only Classifier - Classifier built using both audio and visual features extracted only from the presenter video. Speaker_Audience Classifier - Classifier built using both audio and visual features extracted from both the presenter video and the audience video.

Table 4.1: Speaker Ratings Classifiers

Type	**Name**
Audio-only Classifier	Pitch - max, range
	Intensity - max, range, mean
	Speech Rate, Articulation Rate
Visual-only Classifier	Speaker head movement - mean
	Speaker Gesticulation - mean, std deviation
Audio-Visual Classifier	Pitch - max, range
	Intensity - max, range, mean
	Speech Rate, Articulation Rate
	Speaker head movement - mean
	Speaker gesticulation - mean, std deviation.

Table 4.2 shows the details of the classifiers trained for classification of audience engagement and the features used to train them.

Table 4.2: Audience Engagement Classifiers

Type	**Name**
Audience visual-only Classifier	Audience motion normalised by speaker motion* - mean.
	Audience motion total - mean.
	Frontal face counts - mean.
	Attendance.
Speaker Based Features-only Classifier	Intensity - max, mean, std deviation
	Speaker motion - mean
	Articulation Rate
	Final Speaker Rating
Speaker_Audience Classifier	Intensity - max, mean.
	Articulation rate.
	Speaker Motion - mean.
	Final Speaker Rating
	Audience motion normalised by speaker motion* - mean.
	Audience motion total - mean.
	Frontal face counts - mean.
	Attendance.

* - As the speaker is in the video of the audience, we normalise the total motion to account for that which is motion of the speaker, in order to gain a figure for audience motion only.

4.1.5 Linear Correlations to Presentation Ranking and Engagement

This section examines linear correlations between individual speaker and audience based modalities, and human annotations for speaker ratings and audience engagement levels. Results are shown in two tables, and the modalities with the strongest correlations to human annotations are highlighted.

Table 4.3 shows the linear correlation scores between multimodal speaker-based features and annotated speaker ratings, while Table 4.4 shows linear correlations with annotated audience engagement measures.

Table 4.3: Speaker Ratings - Linear Correlation

Multimodal Feature	$r =$
Pitch Mean	-0.135
Pitch Max	**0.185**
Pitch Range	**0.190**
Pitch Std Dev	-0.044
Intensity Mean	**0.374**
Intensity Max	**0.336**
Intensity Range	**0.218**
Intensity Std Dev	0.141
Speech Rate	**0.313**
Articulation Rate	**0.459**
Face Movement Range	-0.065
Face Movement Mean	**-0.233**
Face Movement Std Dev	-0.106
Face Movement Total	**-0.210**
Speaker Motion Range	**0.302**
Speaker Motion Mean	**0.558**
Speaker Motion Std Dev	**0.415**

Of the modalities showing the strongest correlations with speaker ratings, we can see that Speech Rate and Articulation Rate have a big influence on how strong a presenter's annotators perceived this speaker to be. This can be seen from the bold text figures in Table 4.3. Visual modalities, such as overall speaker movement normalised by head movement, appear also to show a strong correlation with speaker ratings, while head movement by itself does not appear to have a strong effect. The slight negative linear correlation of head movement could suggest that too much head movement may actually put listeners off. Table 4.3 indicates a slight correlation between lower speaker ratings and high levels of head movement.

From Table 4.3, we can see that a number of speaker based acoustic and visual modalities show a strong correlation with audience engagement levels. The most influential of these are clearly the mean and max intensity values. Intensity range and standard deviation values are also effective, however these do not appear to be as important as the max and mean values.

Table 4.4: Engagement Levels Linear Correlation

Multimodal Feature	$r =$
Pitch Mean	0.014
Pitch Max	0.067
Pitch Range	0.062
Pitch Std Dev	0.044
Intensity Mean	**0.318**
Intensity Max	**0.352**
Intensity Range	**0.191**
Intensity Std Dev	**0.194**
Speech Rate	-0.021
Articulation Rate	**0.135**
Face Movement Range	0.040
Face Movement Mean	-0.033
Face Movement Std Dev	0.013
Face Movement Total	-0.044
Speaker Motion Range	-0.004
Speaker Motion Mean	**0.128**
Speaker Motion Std Dev	0.044

Articulation rate is also shown to be an important modality for prediction of audience engagement, as highlighted in Table 4.4. Of the visual modalities the mean speaker motion normalised by head movement is shown to be quite influential. However head movement alone does not appear to be anywhere near as effective as overall speaker motion for prediction of engagement, suggesting that speaker motion should be comprised of overall body motion and gesturing rather than head movement on its own.

Speaker Intensity on the other hand, measured here by loudness, shows a very similar positive linear correlation with both speaker ratings and audience engagement, as can be seen from Table 4.3 and Table 4.4. From these results, we conclude that we can use mean and maximum intensity values as a basis for the prediction of both speaker ratings and audience engagement levels.

4.1.6 Classification of Speaker Ratings

In this section we document results from the classification of speaker ratings. As described in Section 4.1.4.2, we use an ordinal class classifier for classification, evaluating using both 10-fold cross validation and Leave-One-Out cross validation. Classification was performed over an 8-class range and a 4-class range.

Table 4.5: 8-Class Speaker Ratings Classifier - 10FCV

Modalities	Accuracy	MAE	RMSE
Baseline	**30.194**	**1.129**	**1.527**
Audio Only	51.177	0.538	0.643
Visual Only	37.052	0.805	1.214
Audio and Visual	52.098	0.533	0.650

Table 4.6: Audio-visual confusion matrix - 10FCV

	A	B	C	D	E	F	G	H
A	11	15	2	0	0	0	0	0
B	8	37	18	3	0	0	0	0
C	0	10	66	35	5	1	0	0
D	0	4	43	128	83	8	0	0
E	0	1	7	69	178	40	0	0
F	0	0	1	10	54	58	9	1
G	0	0	0	0	5	17	28	6
H	0	0	0	0	1	0	12	3

Table 4.5 shows classification results for speaker ratings over an 8-class range. Tables 4.6, 4.7 and 4.8 show confusion matrices for each modality. From Table 4.6, we can see that Audio-Visual modalities achieve an accuracy of 52.1%, with Audio only modalities achieving an accuracy just below this of 51.1%. Visual-only modalities are shown to result in a less effective classifier, achieving accuracy of just

Table 4.7: Audio-only confusion matrix - 10FCV

	A	B	C	D	E	F	G	H
A	7	20	0	1	0	0	0	0
B	7	29	24	6	0	0	0	0
C	0	3	68	36	10	0	0	0
D	0	0	34	129	94	9	0	0
E	0	0	5	77	182	30	1	0
F	0	0	0	7	64	52	9	1
G	0	0	0	0	5	16	30	5
H	0	0	0	0	1	1	11	3

Table 4.8: Visual-only confusion matrix - 10FCV

	A	B	C	D	E	F	G	H
A	0	17	4	5	2	0	0	0
B	1	20	18	23	4	0	0	0
C	0	9	39	48	21	0	0	0
D	1	9	34	88	126	5	3	0
E	1	4	13	74	181	18	4	0
F	0	0	2	19	95	11	6	0
G	0	0	0	0	23	9	22	2
H	0	0	0	0	1	1	13	1

37%, although the inclusion of visual modalities does slightly increase accuracy over audio-only modalities.

Table 4.9 shows classification results for speaker rating over an 8-class range evaluated using leave one out cross validation. These are again followed by confusion matrices for each modality, shown in Tables 4.10, 4.11 and 4.12. With little change in classification accuracies compared to evaluations with 10-fold cross validation, audio-visual modalities achieve an accuracy of 50.4%, with Audio only modalities achieving an accuracy of 52.3%. Visual only modalities achieve an accuracy of 40%. Results from 10-fold cross validation and leave-one-out cross validation can differ. Leave-one-out CV gives estimates of test error with lower bias and higher variance than 10-fold CV due to the high amount of overlap. The small difference in these results could suggest a high degree of reliability of these results.

Table 4.13 shows classification results for speaker ratings over a 4-class scale. These are again followed by corresponding confusion matrices in Tables 4.18, 4.19

Table 4.9: 8-Class Speaker Ratings Classifier - LOOCV

Modalities	Accuracy	MAE	RMSE
Baseline	**30.194**	**1.129**	**1.527**
Audio Only	52.303	0.529	0.640
Visual Only	40.020	0.779	1.179
Audio and Visual	50.461	0.551	0.673

Table 4.10: Audio-visual confusion matrix - LOOCV

	A	B	C	D	E	F	G	H
A	13	12	2	0	1	0	0	0
B	12	33	20	0	1	0	0	0
C	0	13	58	39	7	0	0	0
D	0	4	35	131	88	8	0	0
E	0	1	10	78	168	38	0	0
F	0	0	1	6	61	59	6	0
G	0	0	0	0	6	15	31	4
H	0	0	0	0	0	2	14	0

Table 4.11: Audio-only confusion matrix - LOOCV

	A	B	C	D	E	F	G	H
A	7	21	0	0	0	0	0	0
B	7	29	33	6	1	0	0	0
C	1	2	69	36	9	0	0	0
D	0	0	29	133	90	14	0	0
E	0	0	1	62	191	39	2	0
F	0	0	1	4	70	47	11	0
G	0	0	0	0	8	11	33	4
H	0	0	0	0	1	0	13	2

Table 4.12: Visual-only confusion matrix - LOOCV

	A	B	C	D	E	F	G	H
A	2	14	4	5	3	0	0	0
B	0	23	28	12	3	0	0	0
C	0	3	60	30	24	0	0	0
D	1	7	36	88	124	9	1	0
E	0	4	16	59	193	19	4	0
F	0	0	2	12	110	4	5	0
G	0	0	0	3	27	4	21	1
H	0	0	0	0	3	0	13	0

and 4.20. As shown in Table 4.13, Audio-visual modalities achieve accuracy of just over 73%, with audio-only modalities achieving accuracy of just over 72%. Once again, Visual-only modalities are weaker, achieving accuracy of just 61%, but can

Table 4.13: 4-Class Speaker Ratings Classifier - 10FCV

Modalities	Accuracy	MAE	RMSE
Baseline	**43.808**	**0.658**	**0.922**
Audio Only	72.364	0.277	0.279
Visual Only	61.208	0.405	0.440
Audio and Visual	73.081	0.272	0.278

Table 4.14: Audio-only confusion matrix - 10FCV

	A	B	C	D
A	64	29	1	0
B	3	271	109	0
C	0	93	328	7
D	0	0	28	44

Table 4.15: Visual-only confusion matrix - 10FCV

	A	B	C	D
A	43	45	6	0
B	18	213	150	2
C	6	108	306	8
D	0	3	33	36

Table 4.16: Audio-visual confusion matrix - 10FCV

	A	B	C	D
A	71	22	1	0
B	14	261	108	0
C	2	83	335	8
D	0	0	25	47

once again be demonstrated to improve accuracy over that for audio-only modalities.

Table 4.17: 4-Class Speaker Ratings Classifier - LOOCV

Modalities	Accuracy	MAE	RMSE
Baseline	**43.808**	**0.658**	**0.922**
Audio Only	**74.514**	**0.256**	**0.258**
Visual Only	62.948	0.386	0.417
Audio and Visual	72.467	0.278	0.285

Table 4.17 shows classification results for speaker ratings over a 4-class scale evaluated with leave-one-out cross validation. Following this, Tables 4.18, 4.19 and 4.20 show corresponding confusion matrices. From Table 4.17, Audio-visual modalities achieve accuracy of 72.5%, with Audio-only modalities achieving accuracy

Table 4.18: Audio-only confusion matrix - LOOCV

	A	B	C	D
A	64	29	1	0
B	3	265	115	0
C	0	68	347	13
D	0	0	20	52

Table 4.19: Visual-only confusion matrix - LOOCV

	A	B	C	D
A	39	49	6	0
B	11	214	156	2
C	4	88	327	9
D	0	3	34	35

Table 4.20: Audio-visual confusion matrix - LOOCV

	A	B	C	D
A	70	22	2	0
B	18	260	105	0
C	1	92	329	6
D	0	0	23	49

of 74.5%. Visual-only modalities achieve an accuracy of just 63%. Once again there is little variation between results of 10-fold CV and leave-one-out CV.

This section documented classification models and results for the classification of a speakers rating. All corresponding confusion matrices were also shown so that full results may be inspected in detail. Highly promising results were achieved, in which we showed the possibility to automatically predict a speaker's rating to accuracies of 52% over a 8-class classification range, and 73% over a 4-class classification range. In the following section we expand upon this and investigate the potential to use these classifiers for the prediction of audience engagement levels during academic presentations.

From Tables 4.5 and 4.13, we can see that the audio features are by far the most important features for classification of speaker ratings. Classification using the visual-only features by comparison achieved results far inferior to those achieved using audio-only features. Classification using both audio and visual based features achieves results only slightly better than those achieved for audio-only, which are not statistically significant. This is also to be expected as the primary source of information is this domain is the audio-stream.

4.1.7 Classification of Audience Engagement

In this section we present results of our investigation into the automatic prediction of audience engagement levels during academic presentations. As described in Section 4.1.4.2, we use an ordinal class classifier for classification, evaluating with both 10-fold cross validation and leave-one-out cross validation. Classification was performed over a 3-class range.

Table 4.21: Audience Engagement Classifier - 10FCV

Modalities	Accuracy	MAE	RMSE
Baseline	**62.436**	**0.376**	**0.615**
Audience Visual Only	66.121	0.342	0.348
Speaker Based Prediction Only	68.270	0.322	0.333
Audience Visual and Speaker	70.317	0.303	0.315

Table 4.22: Speaker-based only confusion matrix - 10FCV

	A	B	C
A	96	92	2
B	52	521	37
C	2	124	50

Table 4.23: Audience visual & Speaker confusion matrix - 10FCV

	A	B	C
A	96	90	4
B	53	519	38
C	1	103	72

Table 4.24: Audience visual-only confusion matrix - 10FCV

	A	B	C
A	79	110	1
B	53	525	32
C	2	132	42

Table 4.21 shows classification results for audience engagement over the 3-class range evaluated with 10-fold cross validation. Following this, Tables 4.22, 4.23

and 4.24 show the corresponding confusion matrices. Engagement can be classified to an accuracy of 70% using speaker and audience based modalities. Using speaker based modalities only we can predict engagement to an accuracy of 68%, while an accuracy of 66% can be achieved by training on Audience based visual modalities only.

Table 4.25: Audience Engagement Classifier - LOOCV

Modalities	Accuracy	MAE	RMSE
Baseline	**62.436**	**0.376**	**0.615**
Audience Visual Only	66.940	0.334	0.340
Speaker Based Prediction Only	66.018	0.345	0.355
Audience Visual and Speaker	72.364	0.282	0.295

Table 4.26: Audience-visual only confusion matrix - LOOCV

	A	B	C
A	78	111	1
B	56	518	36
C	2	116	58

Table 4.27: Audience visual & Speaker confusion matrix - LOOCV

	A	B	C
A	110	77	3
B	59	517	34
C	3	93	80

Table 4.28: Speaker-based only confusion matrix - LOOCV

	A	B	C
A	81	107	2
B	50	502	58
C	3	111	62

Table 4.25 shows classification results for audience engagement over the 3-class range evaluated with leave-one-out cross validation. Following this, Table 4.28, Table 4.27 and Table 4.26 show the corresponding confusion matrices. Engagement can be classified to an accuracy of 72% using speaker and audience based modalities. Using speaker based modalities only we can predict engagement to an accuracy of

66%, while an accuracy of 67% can be achieved by training on Audience based visual modalities only. In addition to mean and maximum intensity, the presenter's articulation rate, mean speaker motion and predicted speaker ratings can be used to predict audience engagement to an accuracy of 68%. By using mean facial counts per second, mean audience motion and extracted attendance rates we are able to improve classification accuracy to 70%. Table 4.21 also shows results for intensity mean, intensity range, intensity, standard deviation and maximum intensity only.

Our results for intensity relate directly to the findings in (Oertel et al., 2011), which shows that increasing levels of intensity correspond to increasing levels of involvement. Our findings on body movement can also be related to this previous work in which they found that increasing levels of body movement corresponding to levels of involvement. This bears some resemblance to our results which show the importance of mean body motion, but not of the standard deviation and range of movement, as seen in Table 4.4.

This section presented classification results for the automatic prediction of audience engagement levels during academic presentations. Classifiers were built using Speaker-based modalities only, Audience-based visual modalities only, and Speaker and Audience-based modalities. An ordinal class classifier was used for classification and evaluated using both 10-fold cross validation and leave-one-out cross validation. We have demonstrated promising results for the prediction of audience engagement.

Emotion features, extracted for classification purposes in the next chapter on comprehension, have not been used in this chapter. This is because we have been able to achieve good classification results for audience engagement just by using the features as described in this chapter. Due to the fact that it was desired to limit the number of features used for classification, and to avoid over-complicating this task, we have not used emotion features in this chapter as these were not considered necessary to achieve good results.

4.1.8 Speaker Ratings : Audience Engagement - Linear Correlations

In this section we compare linear correlations between speaker ratings and audience engagement levels in order to investigate the relationship between these concepts and discover how much of audience engagement can be said to be content based.

Table 4.29: Speaker Ratings : Engagement - Linear Correlation

Label	Label	$r =$
Speaker Ratings	Audience Engagement	**0.227**

The weak positive linear correlation between speaker effectiveness and engagement annotations, as shown in Table 4.29, suggests that there may be a difference between what people popularly perceive to be good public speaking techniques and what is more likely to capture the attention of the audience. We obtain this score by calculating the Pearson's correlation coefficient between human annotations on speaker ratings and audience engagement levels on this dataset, as explained in Section 4.1.4.2. The most obvious of these speaking techniques appears to be articulation rate, which has a strong positive linear correlation with speaker ratings, but a much weaker correlation with audience engagement levels. Similarly the use of gesturing and other speaker movements tended to attract positive speaker ratings, but again shows a much weaker correlation with actual audience engagement. Nevertheless, the r score of 0.227 suggests a linear correlation which can be fused with pre-extracted multimodal speaker-based features to help with prediction of audience engagement levels, and does indicate that engagement is not solely content based either.

In this section we have described the relation between a presenter's speaker ratings and engagement levels during their presentations. We have shown a slight correlation, indicating that a presenter's speaker ratings can be used as a predictive feature for the classification of engagement. In the next part of this chapter, we move on to identifying emphasised speech during these presentations.

4.2 Identification of Intentional or Unintentional Emphasised Speech

In this section, we investigate the identification of emphasised speech, whether intentional or unintentional, within academic presentations. We also investigate potential correlations between instances of emphasised speech and levels of audience engagement and speaker ratings.

4.2.1 Introduction

During a presentation, the presenter can sometimes wish to put extra emphasis on certain points during the talk. This can be to indicate the importance of a specific word or phrase, or alternatively this may indicate the structure of a talk as the speaker moves to a different section. For example, during the spoken phrase "We used an *ordinal class* classifier", with the applied emphasis on "ordinal class", we can deduce that the presenter has emphasised this part as they feel "ordinal class" is a particularly important point for the listener to take note of as distinct from just any classifier. On the other hand, if a presenter emphasises the words "*Results*", and "*Conclusions*", during their presentation, we could infer that the speaker means to identify these words as the beginning of a new section of their talk.

Sometimes however, a presenter may unintentionally apply emphasis on certain points of a presentation through unmanaged use of gesturing, movement or change of tone, in which a listener may note the gesticulation or change of tone and take this as a marker of intended emphasis.

In this part of the chapter we address the following research question:

3. **'Can visual and acoustic stimuli on the part of the speaker be utilised to discover areas of special emphasis being provided by the speaker to indicate important parts of their presentation?'**

Following this research question, if we can detect spoken emphasis, this leads us to address a secondary research question emanating from this:

- Secondary RQ (RQ-3-2): **'Is there a relationship between the concepts of speaker ratings and emphasised speech, and audience engagement and emphasised speech?'**

Previous work on the detection of emphasised speech, as explored in our discussion of related work in Section 2.2.3, has investigated the detection of emphasised speech in the audio-stream only. In this study, we investigate the detection of emphasised speech in the audio-visual domain of academic presentations. The previous section of this chapter focussed on the detection of engagement in audio-visual presentations, in this section, following the identification of emphasised regions of audio-visual presentations, we study potential correlations between emphasised regions of speech and audience engagement at a fine-grained level. To achieve this, as described in Section 3.4.2, the dataset was labelled for levels of engagement on an ordinal scale from 1 to 6, at a fine-grained time scale of 6-seconds.

4.2.2 Multimodal Feature Extraction

For our investigation of emphasis in academic presentations, we extracted the following audio-visual features from the recorded presentations: Pitch, Intensity, Head Movement, and Speaker Movement. These extraction processes used the same methods as described earlier in Section 4.1.3.1 and Section 4.1.3.2.

All features were normalised over their entire range for each speaker and average values calculated over 1-second intervals.

4.2.3 Experimental Investigation

We now describe our experimental investigation into the identification of emphasised parts of audio-visual presentations. The purpose of this investigation was to explore the extent to which we are able to automatically detect areas of intentional or unintentional emphasis being applied by the presenter. The automatic detection of emphasis in audio-visual recordings could potentially allow for the utilisation of this feature in the automatic summarisation of academic presentations.

4.2.3.1 Identification of Emphasis

The first part of this investigation involved asking a total of 10 human annotators, each of whom were research students at Dublin City University, to each watch two of the 5-minute video clips taken from the four presentations. These four videos were chosen for the investigation as all appeared to contain many instances of emphasised speech, which did not appear to occur so frequently in the other videos of the dataset. The annotators were asked to mark areas where they consider the presenter to be giving emphasis. There was actually much disagreement between the annotators over areas of emphasis. To better understand the characteristics of regions consistently labelled as emphasised, we examined the areas of agreed emphasis between the annotators. It was clear from this analysis, that consistent with earlier work, all agreed areas of emphasis occur during areas of high pitch, but also in regions of high visual motion coinciding with an increase in pitch. Following this an extraction algorithm was developed using the following features extracted to locate candidate areas of emphasis: Pitch, Intensity, Head Movement, and Speaker Movement, described earlier in Section 4.2.2.

The algorithm selects candidate regions by finding areas of high pitch occurring in combination with areas of high motion or head movement, as was found to occur during the analysis described above. A two-second gap was allowed between areas of high pitch and high movement on the part of the speaker for selection of areas of emphasis. This two-second gap allows for a short time lag in the detection of these features which can still be said to be occurring together.

Candidate emphasised regions, which were later evaluated by additional human annotators, were marked from extracted areas of pitch within the top 1 percentile of pitch values, top 5 percentile of pitch values and top 20 percentile of pitch values, in addition to top 20 percentile of gesticulation down to the top 40 percentile of values respectively. This resulted in 83 candidate areas of emphasis from our dataset. These candidate regions were each judged for emphasis by three human judges, with the majority vote on each candidate emphasis region taken as the gold standard label for final agreement of emphasis.

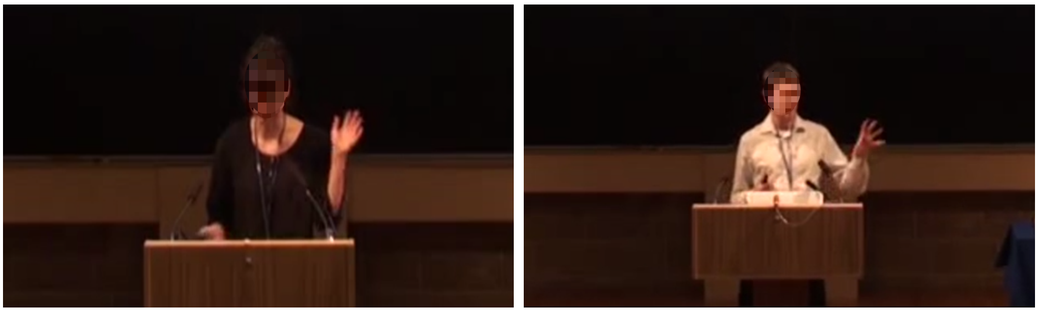


Figure 4.1: Images of speakers emphasising parts of their presentations by using gesturing in addition to increased pitch over specific words or phrases.

4.2.3.2 Correlations with Audience Engagement

The original annotation of this data consisted of annotation of audience engagement over 6-second intervals to measure engagement at a fine grained level. This is over the same data as has now been annotated for regions of emphasis. Despite the fact that initial annotations of emphasis had much disagreement, later annotation by three separate annotators on each candidate area of emphasis gave us final agreement on what constitutes emphasis.

The value of fine-grained annotation of engagement is that this allows us to determine whether intentional or unintentional speech emphasis produced by the presenter correlates with increases or decreases in audience engagement levels at a fine scale. Later, we will look at potential correlations over the larger scale of 30-second speech segments used for engagement annotations, which allows us to also calculate any correlations with speaker ratings, by calculating the Pearson's correlation coefficient between the number of emphasised regions per section and the averaged engagement value for that section.

4.2.4 Results and Analysis

In this section we report the results of our investigations into the identification of emphasised speech in audio-visual recordings. We also introduce results of our investigation in to correlations between emphasised speech and speaker ratings, and between emphasised speech and audience engagement levels.

4.2.4.1 Identification of Emphasised Speech

Of the 83 candidate areas of emphasis extracted from presentation segments, 18 had pitch values in the top one percentile after normalisation. Of these 18 candidate areas, four were accompanied by speaker motion, mostly gesturing, sometimes head movement, while 14 were not accompanied by any speaker movement or gesturing of any significance. All of the 4 candidate areas accompanied by movement or gesturing were judged by human annotators to be emphasised regions of speech. Only 5 of the 14 candidate areas not accompanied by gesturing or movement of any sort were judged by human annotators to be emphasised speech. This indicates that in audio-visual context, emphasised speech frequently depends on gesturing and / or other movement in addition to pitch.

Fifteen of the candidate areas of emphasis were in the top 5 percentile of pitch values extracted. Three of these were accompanied by gesturing on the part of the presenter. All three of these areas accompanied by gesturing were judged by human annotators to be emphasised speech. Of the 12 areas not accompanied by any gesturing by the presenter, only 5 were judged to be emphasised by our human annotators. A total of 33 emphasis candidates were extracted from pitch values in the top 5 percentile. Seven were accompanied by gesturing and all of these were judged by human annotators to be emphasised. Twenty-six were not accompanied by gesturing, and only 10 of these were judged by the human annotators to be emphasised. It was found that candidate emphasis regions in the top 20 percentile of pitch values and the top 20 percentile of gesticulation combined were true regions of emphasis, as labelled by our human annotators. To assess the level of inter-annotator agreement for this task, we calculated the intra-class correlation model 1, ICC(1,1). This was calculated as 0.5818, giving us a good level of inter annotator agreement between judges.

As the examples used thus far provided very few samples to definitively state reliable results, we extracted 15 additional samples of emphasised speech from the corpus. These were extracted from areas where normalised motion and pitch both exceed the top 20 percentile with a two-second gap. In addition, 13 additional samples of non-emphasised speech were used. Three additional human annotators were recruited to annotate new candidate emphasis area. Thirteen of the 15 emphasised areas were labelled by human annotators as emphasised speech.

As indicated by the above results, all annotated areas of emphasis contain significant gesturing in addition to pitch with the top 20 percentile. Gesturing was also found to take place in non-emphasised parts of speech, however this was much more casual and not accompanied by pitch in the top 20 percentile.

4.2.4.2 Correlations Between Speaker Ratings and Emphasised Speech - 30 seconds

We next calculate potential correlations between annotated speaker ratings and annotated emphasised speech. To achieve this we took values for four separate 5-minute video clips containing original emphasis annotations. We achieved this by first calculating the average speaker rating for each 90-second time window, then summing the total number of emphasis detections within that time-frame. Time-windows are incremented at each step by 30 seconds.

Calculating this over the four 5-minute video clips used in this study combined gives a total of 32 time-windowed instances over 20 minutes of video. We calculate correlations using the Pearson's Correlation Coefficient Calculator. Following this, we also calculate the correlation coefficient for speaker specific correlations between speaker ratings and emphasised speech. Table 4.30 outlines the results of these tests.

Table 4.30: Speaker Ratings - Emphasis : Linear Correlation

Video	$r =$
All_Combined	-0.3247
Video 1	-0.2988
Video 2	-0.0845
Video 3	-0.3362
Video 4	0.7976

Although the calculation for all videos combined shows a weak but nonetheless existent negative correlation between speaker ratings and emphasis, when we look at the calculations for all videos we see that Video 4 alone holds a strong positive correlation of 0.7976. With all other videos in the set showing a weak negative correlation, we can conclude that no true correlation exists between speaker ratings and emphasis.

4.2.4.3 Correlations Between Audience Engagement Levels and Emphasised Speech - 6 seconds

We next calculate potential correlations between audience engagement levels as annotated by our human annotators, and instances of emphasised speech, also as annotated by our human annotators. To calculate this we take emphasis annotations over each of the four 5-minute videos used for this study. We first calculate the average engagement level for each 18-second time window, then summing the total number of emphasis detections within that time-frame. Time-windows are incremented at each step by 6 seconds. We use this method to enable us to easily calculate the correlation coefficient simply by aligning average engagement levels per 18-second time window with the number of instances of emphasised speech in that window.

Calculating the correlation coefficient over all of the 5-minute video clips combined gives a total of 188 time-windowed instances. Potential correlations between emphasised speech and audience engagement are calculated using the Pearson's Correlation Coefficient Calculator. Following this, we also calculate the correlation coefficient for speaker specific correlations between audience engagement levels and emphasised speech over 30-second intervals. Table 4.31 shows the results, of which no clear correlation appears between these two concepts.

Table 4.31: Audience Engagement - Emphasis : Linear Correlation

Video	$r =$
All_Combined	-0.0909
Video 1	-0.1449
Video 2	0.1546
Video 3	0.1482
Video 4	0.2108

From Table 4.31, we can clearly see that no real correlation exists, as no clear pattern emerges of correlations per video. Results show very weak positive and negative correlation coefficients, leading us to conclude that there is no correlation between emphasised speech and audience engagement levels. The next subsection looks at possible correlations over 30-second intervals and 90-second time windows, to investigate potential correlations over a more coarse-grained level than already investigated.

4.2.4.4 Correlations Between Audience Engagement Levels and Emphasised Speech - 30-seconds

In this section, we calculate potential correlations between annotated audience engagement levels and annotated emphasised speech. This differs from the previous calculations above by calculating potential correlations over 90-second time windows, to investigate potential correlations occurring over a more coarse-grained level. Once again, to achieve this we take emphasis values for four separate 5-minute video clips containing original emphasis annotations. We first calculate the average engagement level for each 90-second time window, then summing the total number of emphasis detections within that time-frame. Time-windows are incremented at each step by 30-seconds. This enables us to easily calculate the correlation coefficient between engagement levels and instances of emphasised speech.

Calculating this over all of the 5-minute video clips combined gives a total of

32 time-windowed instances. Correlations are calculated using the Pearson's Correlation Coefficient Calculator. Following this, we also calculate the correlation for speaker specific correlations between audience engagement levels and emphasised speech. Table 4.28 shows the results, of which no clear correlation appears between these two concepts.

Table 4.32: Audience Engagement - Emphasis : Linear Correlation

Video	$r =$
All_Combined	-0.1593
Video 1	-0.475
Video 2	0.2887
Video 3	0.8868
Video 4	0.1857

From Table 4.32 we can clearly see that no pattern emerges for the correlation values for each video, leading us to again conclude that no correlations exist between audience engagement levels and emphasised speech. While Video 3 indicates a strong positive correlation, Video 1 indicates a medium negative correlation while other videos show no real correlation. Overall with no clear pattern emerging we can conclude that no correlation exists. However, it should of course be noted that this analysis is performed over a small set of data. We limited these investigations to a small dataset due to the limited number of videos in the full dataset showing instances of emphasised speech, and also due to the limited number of human annotators available, with each video requiring 5 separate annotators.

4.3 Summary

In this chapter we studied the concept of 'good' speaking techniques and their relationship to audience engagement. A set of audio-visual features from the presenter were extracted. We trained a classifier to automatically rate the qualities of a public speaker using extracted multimodal audio-visual features. We investigated correlations between extracted audio-visual features and human annotated speaker ratings to discover the best features for building a classifier to rate the qualities of a public speaker.

We trained a classifier to attempt to classify levels of audience engagement with the talk in progress through the use of multimodal features. Audio-visual features extracted from the presenter along with visual features extracted from the audience were used to train the classifier along with outputted speaker ratings classified using the earlier classifier built in the chapter. Linear correlations were investigated between audience engagement levels and extracted multimodal features to again attempt to find the best features for building a classifier of this nature.

The second part of this chapter looked at emphasised speech, whether intentional or unintentional, in audio-visual presentations. Human annotators were asked to mark points of presentations that they considered to be emphasised. Conditions were found to identify parts of speech considered intentionally or unintentionally emphasised. These were the occurrence of high pitch values (top 20 percentile) accompanied by high levels of speaker motion (top 20 percentile).

In the next chapter we explore the concept of comprehension, and investigate whether it is possible to predict a speaker's potential to be comprehended by their audience.

Chapter 5

Classification of Comprehensible Material

In this chapter we describe our investigations into the concept of comprehensibility, in which we consider the question of how much the audience can follow and understand the material being presented in an academic presentation. We discuss possible causes for comprehensibility, or lack there of. We investigate the modalities which could provide information as to the comprehensibility of a presentation by an audience. We perform a series of experiments on classification of comprehension, attempting to accurately classify this concept over different classification ranges. We also explore the potential relationship between the concepts of audience engagement and comprehensibility.

5.1 Introduction

Many of us have been present at talks and presentations where the audience is interested in the material and engaged with the presentation, but where large numbers of attendees are still unable to completely follow the content for various reasons. This can sometimes be down to not having the required background knowledge of the subject at hand in order to completely follow. However, other times this can be the result of how the presentation is delivered.

Sometimes the speaker can be delivering information at a rate which is too fast or too slow, other times the speaker's accent can be difficult to understand, or the presenter may speak in a voice which is just too quiet for the audience to hear fully. Often comprehension may be hindered by the layout of the presentation slides not being conducive to understanding of the material. There are also times when the presenter is just unable to entice real interest in the material on the part of the audience, naturally leading to the audience being unable to fully comprehend the material.

On other occasions, the audience finds the presenter is particularly skilled at making the material both interesting and comprehensible. On these occasions, the presenter seems to be capable of making the material comprehensible to anyone even remotely interested in the topic being presented just by explaining it at the right pace, at the right level, with clear explanatory slides, and through speaking clearly.

Having addressed the question of engagement or interest during academic presentations in the previous chapter, in this chapter we look at the concept of comprehension - how much the audience can follow and understand the content. We look at a number of input modalities, including the presentation slides, audio-visual features of the speaker, and visual features from the audience.

In this chapter we address the following research question:

4. **'Is it possible to build a classifier to identify parts of a presentation considered to be most understandable and comprehensible for the user?'**

   Following this research question, if we can classify levels of audience comprehension, this leads us to address a secondary research question relating to this topic:

   - Secondary RQ (RQ-4-2): **'Is there a relationship between the concepts of audience engagement and audience comprehension?'**

We consider audience comprehension to be aided by a number of factors, including, but not limited to: the participant's prior knowledge in the area of the presentation, how interested they are in the topic, the speed and clearness with which the presenter is speaking, the participant's familiarity with the presenters accent, how clear the presenter's slides are, how engaging the presenter makes the presentation, and the absence of sources of distraction for the audience.

We hypothesise that by extracting a number of audio-visual features including visual features of the presentation slides, and additional audio-based features, such as fluency measures and accentuation, we can build a classifier to predict the speaker's potential to be comprehended. While we cannot account for the participant's prior knowledge or their familiarity with the presenter's accent, we can account for the presenter's speaking techniques, how interesting they make the presentation and the absence of sources of distraction, in order to predict the speaker's potential to be comprehended by the audience.

5.2 Objectives

The objectives of this chapter are to investigate the potential of statistical classifiers to predict a speaker's potential to be comprehended by their audience during an academic presentation. Following this, we investigate whether there is a relationship, and the nature of any such relationship, between the concepts of audience engagement and audience comprehension. First, however, we extract individual audio-visual features and calculate the Pearson's Correlation Coefficient between extracted audio-visual features and human annotated scores for comprehension levels. For this we extract many of the same audio-visual features as for the previous chapter, however, we also extract additional features for comprehension which are inspired by our review of related work on comprehension, such as slide cluster and fluency measures. This helps us to identify the most important individual modalities for predicting the speaker's potential to be comprehended by the audience.

To aid us in achieving these objectives, we hypothesise that the type of features we can successfully extract which may be of use in indicating or predicting audience comprehension. In addition to the features from the previous chapter of pitch, intensity, speech rate, articulation rate, the speaker's head movement, speaker motion, general audience motion and audience facial count, for prediction of comprehension we also extract additional audio features to calculate features of speech fluency in addition to speech formants to garner accentual information. A set of emotion features were extracted from audio of the presenter using the OpenSmile emotion dataset (Eyben et al., 2013).

We hypothesise that a clear set of readable slides which are not overly cluttered, in additional to clear, fluent speech, can aid audience members in comprehending the material. By gathering information on the presentation slides we hope to use this information to guide a classifier to make inferences as to the comprehension of the audience. From the review in 2.2.4, the design presentation slides (Garner and Alley, 2013) can have quite an effect on the comprehension of the audience. We use optical character recognition to garner information on the presentation slides, including average word count per slide, average number of lines per slide and the total amount of cluster per slide, or how full with text or images each slide is.

Given the number of different modalities from which we extract features in seeking to classify the likely comprehension of a presentation, we hypothesise that the use of late fusion techniques may improve classification performance over early fusion techniques. This is because the presenter's slides, the audio track and the audience visual are completely different modalities, each providing separate clues as to the comprehensibility status. We investigate this by performing classification over a number of class ranges, using both early and late fusion techniques, and comparing the results achieved.

Finally, we calculate the Pearson's Correlation Coefficient for the linear correlation between the concepts of audience engagement and comprehension to investigate whether there is a relationship between them. At first glance these concepts would appear to have an important relationship, as to fully comprehend a topic one must first be interested in that topic. By aligning human annotations for audience engagement, obtained from the human annotations as described in previous chapters, with annotations for audience comprehension, obtained from crowdsourced human annotations described in Section 3.4.3, and calculating the Pearson's Correlation Coefficient, we can calculate the linear correlation between audience engagement and comprehension, giving us the actual relationship between these two concepts.

The following section describes the extraction of the audio-visual features used for these investigations.

5.3 Multimodal Feature Extraction

Based on our review of previous work in the field of psychology in Section 2.2.4, we conclude that audience comprehension of spoken content is based largely on the fluency of the speech of the presenter in addition to a clear layout of presentation slides (Tanenhaus et al., 1995; Garner and Alley, 2013; Haake et al., 2014).

We propose that audience comprehension in academic presentations is influenced by multiple modality streams, including the fluency and the accentuation of the presenter's speech, layout of the presentation slides and the overall visual information stream available to the attendee, including gesturing and body movement on the part of the speaker.

Based on this analysis we suggest that the best features to extract for our investigation of audience comprehension to be fluency measures of the spoken content, in addition to information relating to the layout of the presentation's slides. In addition to these features, we also extract basic visual information of the presenter, namely head movement and speaker motion, as we consider that similar to audience engagement, certain movements and gesticulation on the part of the speaker may aid understanding for an audience. Basic visual features from the audience to the presentation are also extracted to discover if these features aid classification of comprehension.

The extracted acoustic and visual features were used to train a comprehension classifier. Acoustic features were extracted using Praat (Boersma et al., 2002), see Appendix C.1.3, while visual features were extracted using OpenCV (Bradski and Kaehler, 2008), Appendix C.1.1. An addition set of acoustic features, normally used for emotion detection, was extracted using OpenSmile (Eyben et al., 2013). Visual features for textual layout of presentation slides were extracted using GOCR open-source optical character recognition (Schulenburg, 2010), a tool for optical character recognition which outputs text appearing on slides which can be used to generate statistics on the amount of text per slide.

The extraction of these individual feature sets are described in the following subsections. Based on the work cited above, we consider the most important of these to be the presenter's audio stream and the slides visual stream, the prime modalities from which the audience receives information, which are supplemented by visual features from the speaker, the secondary modality from which the audience receives information, and audience visual streams which can provide information as to the movement of the audience.

5.3.1 Slides - Visual Features

We now describe features extracted from the slides for the training of a classifier.

- **Clutter** values representing how much each slide is filled with text and images were extracted from slides by comparing lighter pixel colours with darker pixel colours and providing a clutter percentage rate. To achieve this, we developed a short program in OpenCV which focussed in on each slide, and RGB pixels were converted to greyscale. Pixels with a greyscale value of less than 128 were taken as black, while pixels with a value of 128 or greater were taken as white. The colour with the smaller number of pixels was taken as a percentage of the other colour.

- **OCR on presentation slides** Optical Character Recognition was performed on each slide using GOCR open-source optical character recognition. The output of this was a text file for each slide containing all words recognised for the slide. The outputted text also contained basic layout information from each slide such as the positioning of text and number of lines of text. From each text file the average character count, average word count, average number of characters per word and average number of lines per slide were extracted for each presentation in the dataset.

5.3.2 Speaker-Based Acoustic Features

Pitch, Intensity, Speech Rate and Articulation Rate were extracted from the audio stream using the same tools as for the investigation of engagement in Chapter 4, see Section 4.1.3.1. In addition, the following extra features were extracted from the audio stream for our investigation of comprehension.

- **Fluency Measures** were calculated using features extracted from the audio stream. These include the mean number of runs (a run being the time between pauses), mean run duration (length of time between pauses), mean pause duration and average number of pauses per minute. These features serve to indicate the fluency with which the presentation was delivered, giving us potentially more useful features for comprehension detection than using only Speech Rate, Articulation Rate and Syllables per Duration, used for the prediction of audience engagement investigated in Chapter 4. Since we hypothesise that comprehension depends much more on the fluency and clearness of speech of the presenter than audience engagement does, which we hypothesise is more effected by gesticulation and speech pitch and intensity, we consider that audience comprehension is affected in a significant way by the fluency of the speaker.

- **Speech Formants** were extracted using Praat. Taking the average, range, variance and standard deviation of F1, F2, F3 and F4 formant values from across the entire speech segment, in this case the uninterrupted audio-visual clips into which each presentation was segmented for classification of comprehension, we also train on F1 values, F2-F1 values, F3-F4 values, F3-(F2-F1) values and finally F4/F1 values. Often the first two formants are enough to disambiguate the vowel. As we wish to study comprehension here, we therefore took everything into consideration which we suspected could ultimately aid in this, we took all the formats F1 to F4 into consideration and used all of these formants for gaining accentuation information which we expected could influence the comprehensibility of a speech segment for training our classifiers.

5.3.3 Emotion Features

Emotion Features were extracted from the audio content using OpenSmile (Eyben et al., 2013). The set of 384 emotion features from the 2009 Emotion challenge feature set (Schuller et al., 2009) were extracted. We reduced this to the 20 best performing features for classification to avoid the use of features least likely to aid in anyway the classification of comprehension, and to reduce it to a manageable set of features. This set was reduced to the best performing features using the Chi Squared Attribute Evaluator and Ranker Search Method. Further evaluation on this found that the 20 retained features gave better accuracy on classification of comprehension than the full set of 384 features. As the speech features of the emotion dataset had already been found to aid in the classification of emotional speech (Schuller et al., 2009), we sought to investigate whether the same features could also be of use for classification of audience comprehension, due to the role we considered the speakers fluency and accentuation played in comprehension.

5.3.4 Speaker-Based Visual Features

Head Movement and Speaker Motion were extracted from the speaker visual stream, as described in Section 4.1.3.2. We calculate the mean and standard deviation of head movement and of motion, which gives a general value on the speakers gesticulation and movement. We calculate the mean Head movement per second and the mean motion per second, this also applies for standard deviation values. These features were extracted as we consider that certain movements in the part of the speaker may aid in understanding for an audience.

5.3.5 Audience-Based Visual Features

Audience Facial Counts and Audience Motion were extracted from the audience visual stream, as described in Section 4.1.3.3. As for above, we calculate the mean and the standard deviation of audience motion per second, and of frontal facial counts per second. This a secondary stream which can aid in classification of audience comprehension by providing information on the movement of the audience and estimates of numbers of attendees facing towards the presentation.

Having described the multimodal feature extraction process and the reasoning behind the extraction of such features, in the following section we describe the experimental investigations performed to order to address the research questions outlined in the introduction to this chapter.

5.4 Experiment Design

As with the investigation described in the previous chapter, while most similar work in this area has also used a regression model for classification, we use an Ordinal Class Classifier for classification over the multiple classes. Regression models are based on the assumption that all classes are equally spaced from one another, with the difference between the bottom two classes being the same as the difference between the top two classes. However, for this work we know that this is not the case since classes are rated on an ordinal scale, not a regression scale, we have no way of telling whether the difference between the bottom two classes is the same at that between the top two classes.

Previous work on this topic is described in our review of related work in Section 2.2.4. As described above, extracted audio-visual features are based on findings from previous work on audience comprehension. Upon extraction of multimodal features, training sets for training of a classifier were developed using the values calculated as described above from these audio-visual features, separated by modalities, and combined with averaged groundtruth values as derived from crowdsourced human annotations of this concept, as described in Section 3.4.3.

As described in more detail earlier in Section 3.4.3, annotations were made for video segments each between two and four minutes in length. Segment boundaries were placed at changes of topic within talks. Three annotations per video segment were sought. Annotation was performed using the crowdsourcing website *Prolific Academic*[1]. While annotations were made over a base 8-class scale, again to encourage annotators to make a decision rather than choosing the middle option, this was reduced to a 7-class scale since no annotations were made on the highest class of comprehension. Following classification experiments over the 7-class range, the dataset was later combined to form a 4-class range, then further combined to form a binary range. Classes were combined and classification performed over these combined classes to investigate whether the concept of comprehension could be seen as a simple binary classification of comprehensible or incomprehensible speech, in addition to levels of comprehension as originally investigated.

5.4.1 Experimental Investigation

In this section we describe the experiments we performed examining the classification of comprehension of presentations, and relating this concept to speaker ratings and audience engagement. We first describe the classifier we used for classification

[1] https://www.prolific.ac/

and our reasons for using a classifier of this type. Following this we describe our experiments performed in order to best classify areas of high comprehension levels among the audience to that presentation. We then describe the experiments performed to investigate the relationship between previously provided speaker ratings and audience engagement levels.

The Pearson's Correlation Coefficient was calculated between extracted multimodal features and final averaged audience comprehension levels for every segment of the video dataset to calculate the linear correlations between comprehension and individual features. Datasets were divided between separate modalities of speaker audio, emotion, speaker visual, presentation slides and audience visual. This was to allow for the study of late-fusion and early-fusion techniques of the different information streams associated with the modalities.

5.4.2 Detailed Technical Description

Since an 8-class ordinal scale was used for the human annotation of audience comprehension levels, we use an Ordinal Class Classifier for classification. We investigate the use of early-fusion and late-fusion strategies in our classifiers, taking into account that we expect comprehension to be influenced by our multiple modality streams.

We trained a classifier to automatically predict how comprehensible each video segment is by using Weka data mining workbench (Frank et al., 2010), described in further detail in Appendix C.1.4. Classification was performed using the Ordinal Class Classifier (Frank and Hall, 2001) and evaluated using 10-fold cross validation. Further evaluation was performed using Leave-One-Out cross validation.

We now list all classifiers trained and the features used to train them:

1. Audio-only Classifier

    - Pitch - range, mean, variance & standard deviation.
    - Intensity - range, mean, variance & standard deviation.

- Speech Rate, Articulation Rate, Average Syllable Duration, Pauses per Minute.
- Formants - F1, F4/F1.

2. Audio & Speaker Visual Classifier
   - Pitch - range, mean, variance & standard deviation.
   - Intensity - range, mean, variance & standard deviation.
   - Speaker Motion - range, mean, variance & standard deviation.
   - Speaker Facial Movement - range, mean, variance & standard deviation.
   - Speech Rate, Articulation Rate, Average Syllable Duration, Pause Count, Pauses per Minute.
   - Formants - F1, F2-F1, F3-F4, F3-(F2-F1), F4/F1.
3. Audio & Audience Visual Classifier
   - Pitch range - mean, variance & standard deviation.
   - Intensity - range, mean, variance & standard deviation.
   - Speech Rate, Articulation Rate, Average Syllable Duration, Pauses per Minute.
   - Formants - F1, F2-F1, F3-F4, F3-(F2-F1), F4/F1.
   - Audience Motion - mean, range, variance, standard deviation.
   - Front Facing Audience Faces - mean, variance, standard deviation.
4. Audio & Slides Classifier
   - Pitch - range, mean, variance & standard deviation.
   - Intensity - range, mean, variance & standard deviation.
   - Speech Rate, Articulation Rate, Average Syllable Duration, Pause Count, Pauses per Minute.

   - Formants - F1, F4/F1.
   - Slides - average word length.
   - Slides Clutter - range, mean, variance, standard deviation.

5. Slides-only Classifier

   - Slides - average words, average lines, average word length.
   - Slides Clutter - range, mean, variance, standard deviation.

6. Emotion Features-only Classifier

   - MFCC - sma_[11]_linregc2, sma[3]_skewness, sma[11]_amean, sma[12]_skewness, sma_de[3]_skewness, sma[2]_min, sma[10]_skewness, sma_de[2]_stddev,
   - sma_de[2]_linregerrQ, sma[2]_amean, sma_de[12]_stddev, sma_de[12]_linregerrQ, sma[3]_min, sma[7]_amean, sma[10]_min, sma_de[7]_linregc2, sma[12]_stddev.
   - F0 - sma_de_max.
   - PCM - zcr_sma_de_linregerrQ, zcr_sma_de_stddev.

7. Emotion & Speaker Visual Classifier

   - Speaker Motion - range, mean, variance & standard deviation.
   - Speaker Facial Movement - range, mean, variance & standard deviation.
   - MFCC - sma_[11]_linregc2, sma[3]_skewness, sma[11]_amean, sma[12]_skewness, sma_de[3]_skewness, sma[2]_min, sma[10]_skewness, sma_de[2]_stddev,
   - sma_de[2]_linregerrQ, sma[2]_amean, sma_de[12]_stddev, sma_de[12]_linregerrQ, sma[3]_min, sma[7]_amean, sma[10]_min, sma_de[7]_linregc2, sma[12]_stddev.
   - F0 - sma_de_max.
   - PCM - zcr_sma_de_linregerrQ, zcr_sma_de_stddev.

8. Emotion & Audience Visual Classifier

   - Audience Motion - mean, range, variance, standard deviation.

- Front Facing Audience Faces - mean, variance, standard deviation.
- MFCC - sma_[11]_linregc2, sma[3]_skewness, sma[11]_amean, sma[12]_skewness, sma_de[3]_skewness, sma[2]_min, sma[10]_skewness, sma_de[2]_stddev,
- sma_de[2]_linregerrQ, sma[2]_amean, sma_de[12]_stddev, sma_de[12]_linregerrQ, sma[3]_min, sma[7]_amean, sma[10]_min, sma_de[7]_linregc2, sma[12]_stddev.
- F0 - sma_de_max.
- PCM - zcr_sma_de_linregerrQ, zcr_sma_de_stddev.

9. Emotion & Slides Classifier

    - Slides - average words, average lines, average word length.
    - Slides Clutter - range, mean, variance, standard deviation.
    - MFCC - sma_[11]_linregc2, sma[3]_skewness, sma[11]_amean, sma[12]_skewness, sma_de[3]_skewness, sma[2]_min, sma[10]_skewness, sma_de[2]_stddev,
    - sma_de[2]_linregerrQ, sma[2]_amean, sma_de[12]_stddev, sma_de[12]_linregerrQ, sma[3]_min, sma[7]_amean, sma[10]_min, sma_de[7]_linregc2, sma[12]_stddev.
    - F0 - sma_de_max.
    - PCM - zcr_sma_de_linregerrQ, zcr_sma_de_stddev.

10. All Features Classifier

    - Pitch - min, max, range, mean, variance & standard deviation.
    - Intensity - min, max, range, mean, variance & standard deviation.
    - Speaker Motion - range, mean, variance & standard deviation.
    - Speaker Facial Movement - range, mean, variance & standard deviation.
    - Speech Rate, Articulation Rate, Average Syllable Duration, Pause Count, Pauses per Minute.
    - Formants - F1, F2-F1, F3-F4, F3-(F2-F1), F4/F1.
    - Slides - average words, average lines, average word length.

- Slides Clutter - range, mean, variance, standard deviation.
- Audience Motion - mean, range, variance, standard deviation.
- Front Facing Audience Faces - mean, variance, standard deviation.
- MFCC - sma_[11]_linregc2, sma[3]_skewness, sma[11]_amean, sma[12]_skewness, sma_de[3]_skewness, sma[2]_min, sma[10]_skewness, sma_de[2]_stddev,
- sma_de[2]_linregerrQ, sma[2]_amean, sma_de[12]_stddev, sma_de[12]_linregerrQ, sma[3]_min, sma[7]_amean, sma[10]_min, sma_de[7]_linregc2, sma[12]_stddev.
- F0 - sma_de_max.
- PCM - zcr_sma_de_linregerrQ, zcr_sma_de_stddev.

The above classifiers were trained using all of the listed features. There were 172 instances of audio-visual presentations in our dataset in total to be classified for comprehension. Classifiers were trained on an Ordinal Class Classifier on the full training set model, evaluated using 10-fold cross validation. Further evaluation, with results and confusion matrices also listed in this chapter, was performed using Leave-One-Out cross validation.

5.5 Experimental Results

In this section we describe the spread of annotations among each class of comprehension. We then show our results achieved from the classification of comprehension levels during presentations, discuss the significance of results and show how comprehension in this context relates to the concepts of speaking techniques and audience engagement.

5.5.1 Balance of Data Set and Baseline

Final manual annotations of our data set showed an even spread among the second, third and fourth classes of comprehension, with a reduced spread among the fifth, sixth and seventh class. No annotations were received on the highest, eight grade of comprehension, while only 2 scored in the lowest grade of comprehension levels. Full details are shown in Figure 5.1.

We first trained a classifier to predict audience comprehension levels over the 7-class scale. Following this we combined the first and second classes, and fifth, sixth and seventh classes, to form a 4-class range for classification tasks. Following this we further combine classes to form a binary classification problem in order to demonstrate the effectiveness of this technique at different granularity of classification on audience comprehension.

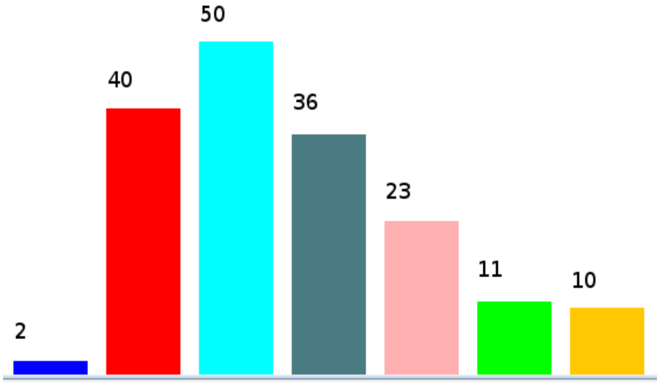


Figure 5.1: 7-Class Distribution

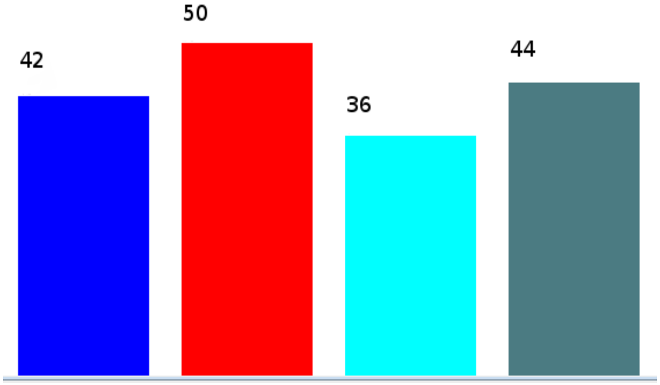


Figure 5.2: 4-Class Distribution

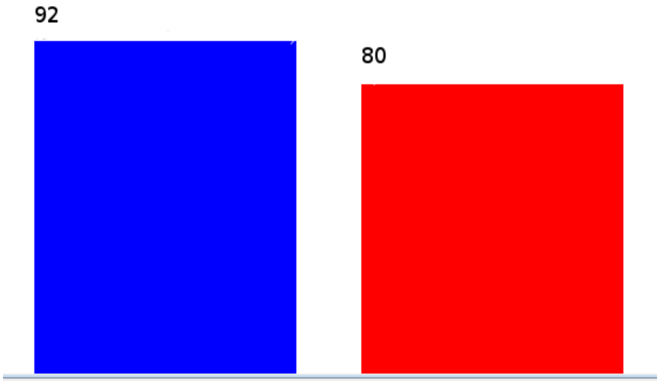


Figure 5.3: Binary Distribution

Figures 5.1, 5.2 and 5.3 show the spread of annotations over each class for each classification scheme: 7-class, 4-class and binary classification. In the absence of any well-defined baselines for work of this nature, we used Zero Rule classification, which simply predicts to the majority class as a baseline to demonstrate the effectiveness of this technique. This is to take account of imbalance in the dataset which gives the results more context.

5.5.2 Early-Fusion Classification Results

This section shows results of early-fusion classification over 7-class, 4-class and binary classification schemes. These are evaluated using both 10-fold cross validation (10FCV) and leave-one-out cross validation (LOOCV). We first show results for a 7-class range evaluated using 10-fold cross validation.

Table 5.1: Early-Fusion Comprehension Classification Results over 7 class distribution - 10FCV

c	Modalities	Acc	MAE	RMSE
-	**Majority Baseline**	**29.070**	**1.157**	**2.529**
1	Audio Only	**45.349**	**0.860**	**1.709**
2	Audio & Speaker Vis	**49.419**	**0.802**	**1.663**
3	Audio & Audience Vis	41.861	0.948	2.052
4	Audio & Slides	**46.512**	**0.808**	**1.576**
5	Slides Only	34.302	1.041	2.041
6	Emotion Features Only	38.372	0.907	1.628
7	Emotion & Speak Visual	43.023	0.837	1.500
8	Emotion & Audience Visual	41.861	0.820	1.413
9	Emotion & Slides	40.116	0.890	1.797
10	All Features	42.442	0.797	1.378

Table 5.1 shows 7-class classification results for all early-fusion classifiers, trained on an ordinal class classifier and evaluated using 10-fold cross validation. Full details including the full list of features used for the training of each classifier are given in Section 5.4.2 above.

Audio & Speaker Visual are shown to have the highest accuracy of early-fusion classifiers, attaining accuracy of 49.4%. Next highest accuracy for early-fusion classifiers is Audio & Slides features with an accuracy of 46.5%, followed by Audio-only with an accuracy of 45.3%. Confusion matrices follow for each classifier. These give a clearer idea of the type of errors each classification model is making.

Table 5.2: Audio-only confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	0	21	9	8	0	2	0
C	0	9	28	9	3	0	1
D	0	4	9	16	5	2	0
E	0	1	3	7	11	0	1
F	0	2	0	5	2	2	0
G	0	0	1	2	4	3	0

Table 5.3: Audio-speaker visual confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	0	21	9	7	2	0	1
C	0	12	25	6	5	2	0
D	0	4	7	19	6	0	0
E	0	2	1	6	14	0	0
F	0	1	1	3	3	2	1
G	0	0	2	1	1	2	4

Table 5.4: Audio-Audience visual confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	0	17	10	7	3	2	1
C	0	9	26	10	3	1	1
D	0	6	10	14	5	1	0
E	0	1	1	8	12	1	0
F	0	2	3	2	1	3	0
G	0	0	1	0	4	5	0

Table 5.5: Audio-Slides confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	0	1	1	0	0	0
B	0	21	12	6	1	0	0
C	0	10	32	5	3	0	0
D	0	4	9	18	4	1	0
E	0	2	3	11	5	2	0
F	0	2	1	2	3	3	0
G	0	1	0	2	3	3	1

Table 5.6: Slides-only confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	0	2	0	0	0	0
B	0	0	25	13	2	0	0
C	0	0	39	8	2	1	0
D	0	0	22	12	1	1	0
E	0	0	15	4	4	0	0
F	0	0	4	0	3	4	0
G	0	0	3	4	2	1	0

Table 5.7: Emotion features-only confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	0	0	2	0	0	0
B	0	13	15	6	6	0	0
C	1	5	28	10	5	1	0
D	0	1	9	12	11	3	0
E	0	0	3	7	10	3	0
F	0	0	1	4	6	0	0
G	0	0	0	2	3	2	3

Table 5.8: Emotion Speaker-visual confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	0	1	1	0	0	0
B	0	12	17	7	4	0	0
C	1	5	31	8	3	1	1
D	0	1	8	14	10	3	0
E	0	0	4	6	11	2	0
F	0	0	0	5	5	1	0
G	0	0	0	2	2	1	5

Table 5.9: Emotion Audience-visual confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	0	0	2	0	0	0
B	0	13	14	6	7	0	0
C	1	7	29	11	2	0	0
D	0	0	12	15	7	2	0
E	0	0	4	7	11	1	0
F	0	0	0	5	5	1	0
G	0	0	0	1	1	5	3

Table 5.10: Emotion + Slides confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	1	0	1	0	0	0
B	0	17	14	5	4	0	0
C	1	10	26	11	2	0	0
D	0	2	12	12	8	2	0
E	0	1	4	7	9	2	0
F	0	1	1	1	3	3	2
G	0	2	2	0	2	2	2

Table 5.11: All features confusion matrix - 10FCV

	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	0	18	16	6	0	0	0
C	0	17	21	8	2	0	2
D	1	3	11	16	2	3	0
E	0	1	5	7	7	7	0
F	0	0	1	3	2	4	1
G	0	0	1	0	0	2	7

From the above confusion matrices we can see that our classifier consistently misclassified instances which had been manually annotated as belonging to the first class of comprehension. However, as this class only contains 2 instances out of a total of 172, we consider that we can safely combine this with the second class of comprehension. Further analysis shows rather poor accuracy with respect to the final three classes of comprehension. Once again however these classes are much smaller than the preceding classes of comprehension in our dataset, as seen in Figure 5.1. We later therefore combine these three classes, giving a much more balanced dataset, as shown in Figure 5.2.

We now show classification results evaluated using leave-one-out-cross validation.

Table 5.12: Early-Fusion Comprehension Classification Results over 7 class distribution - LOOCV

c	Modalities	Acc	MAE	RMSE
-	**Majority Baseline**	**29.070**	**1.157**	**2.529**
1	Audio Only	**45.930**	**0.860**	**1.860**
2	Audio & Speaker Vis	**43.605**	**0.855**	**1.762**
3	Audio & Audience Vis	38.953	1.000	2.233
4	Audio & Slides	**43.023**	**0.884**	**1.733**
5	Slides Only	37.791	0.953	1.930
6	Emotion Features Only	41.279	0.855	1.541
7	Emotion & Speak Visual	41.860	0.884	1.628
8	Emotion & Audience Visual	33.721	0.936	1.634
9	Emotion & Slides	34.884	0.924	1.750
10	All Features	33.721	0.919	1.535

Table 5.12 shows 7-class classification results for all early-fusion classifiers, trained on an ordinal class classifier and evaluated using leave-one-out cross validation. Full details including the full list of features used for the training of each classifier is listed in Section 5.4.2 above.

Audio-only features have shown to give the highest accuracy of early-fusion classification, attaining accuracy of 45.9%. Next highest accuracy for early-fusion classifiers is audio & speaker visual features with an accuracy of 43.6% followed by audio & slides with an accuracy of 43.0%. Once again confusion matrices follow for each classifier. These give a clearer idea of the type of errors each classification model is making.

Table 5.13: Audio-only confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	2	0	0	0	0	0
B	0	15	18	4	2	1	0
C	0	7	35	6	2	0	0
D	0	4	6	19	7	2	0
E	0	4	2	6	9	2	0
F	0	3	1	4	2	1	0
G	0	2	0	0	5	3	0

Table 5.14: Audio-speaker visual confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	0	20	14	5	0	0	1
C	0	7	27	10	3	1	2
D	0	2	13	15	6	0	0
E	0	2	3	8	8	1	1
F	0	1	1	4	4	0	1
G	0	1	1	0	3	0	5

Table 5.15: Audio-Audience visual confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	2	0	0	0	0	0
B	0	17	14	6	2	1	0
C	0	8	29	10	3	0	0
D	0	7	6	13	8	1	1
E	0	3	2	9	5	4	0
F	0	3	3	1	2	2	0
G	0	3	0	0	5	1	1

Table 5.16: Audio-Slides confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	0	17	14	7	0	2	0
C	0	12	28	7	2	1	0
D	0	6	7	15	5	2	1
E	0	2	1	9	10	1	0
F	0	2	1	2	1	4	1
G	0	0	0	1	7	2	0

Table 5.17: Slides-only confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	0	1	0	1	0	0
B	0	0	34	4	2	0	0
C	0	0	47	2	0	1	0
D	0	0	26	9	1	0	0
E	0	0	15	5	3	0	0
F	0	0	4	0	1	6	0
G	0	0	5	2	1	2	0

Table 5.18: Emotion features-only confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	0	2	0	0	0	0
B	0	11	19	5	3	2	0
C	1	2	32	7	7	1	0
D	0	0	13	13	7	3	0
E	0	0	3	10	10	0	0
F	0	0	1	4	3	2	1
G	0	0	0	2	1	4	3

Table 5.19: Emotion Speaker-visual confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	0	1	1	0	0	0
B	0	11	19	8	2	0	0
C	1	3	33	3	6	3	1
D	0	0	12	13	7	4	0
E	0	0	4	9	9	1	0
F	0	0	1	4	5	1	0
G	0	0	0	2	3	0	5

Table 5.20: Emotion + Slides confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	0	2	0	0	0	0
B	0	14	18	5	3	0	0
C	1	11	24	13	1	0	0
D	0	4	14	11	5	2	0
E	0	2	1	10	7	3	0
F	0	1	0	2	2	4	2
G	0	2	1	1	3	3	0

Table 5.21: All features confusion matrix - LOOCV

	A	B	C	D	E	F	G
A	0	0	2	0	0	0	0
B	0	19	15	5	1	0	0
C	0	18	19	9	3	0	1
D	1	3	10	11	7	4	0
E	0	2	5	12	1	3	0
F	0	0	1	3	1	5	1
G	0	0	0	1	4	2	3

As for the previous 10-fold cross validation evaluations, we can see consistent misclassifications for the first class of comprehension as well as poor accuracies over the final three classes of comprehension. We now combine the first class of comprehension with the second, and combine the final three classes, to give a more balanced 4-class dataset, shown in Figure 5.2.

We now show results for early fusion classification performed over the combined 4-class range. Classification is again performed using an ordinal class classifier and evaluated using 10-fold cross validation.

Table 5.22: Early-Fusion Comprehension Classification Results over 4 class distribution - 10FCV

c	Modalities	Acc	MAE	RMSE
-	**Majority Baseline**	**29.070**	**0.965**	**1.477**
1	Audio Only	**52.326**	**0.663**	**1.105**
2	Audio & Speaker Visual	**54.651**	**0.576**	**0.820**
3	Audio & Audience Visual	45.349	0.727	1.157
4	Audio & Slides	**48.256**	**0.703**	**1.145**
5	Slides Only	35.465	0.849	1.291
6	Emotion Features Only	47.093	0.0.721	1.198
7	Emotion & Speak Visual	40.116	0.773	1.180
8	Emotion & Audience Visual	36.628	0.872	1.453
9	Emotion & Slides	41.861	0.797	1.285
10	All Features	51.163	0.581	0.779

Table 5.22 shows 4-class classification results for early-fusion classifiers, trained on an ordinal class classifier and evaluated using 10-fold cross validation. Full details including the full list of features used for the training of each classifier is listed in Section 5.4.2.

Audio & Speaker Visual features are shown to have the highest accuracy of early-fusion classifiers, attaining accuracy of 54.6%. Next highest accuracy for early-fusion classifiers is Audio-only features with an accuracy of 52.3% followed by Audio & Slides with an accuracy of 48.2%. Confusion matrices now follow for each classifier. These give a clearer idea of the type of errors each classification model is making.

Table 5.23: Audio-only confusion matrix - 10FCV

	A	B	C	D
A	20	15	5	2
B	7	31	11	1
C	6	7	17	6
D	4	8	10	22

Table 5.24: Audio-Speaker Visual confusion matrix - 10FCV

	A	B	C	D
A	21	13	5	3
B	8	31	7	4
C	6	7	14	9
D	4	0	12	28

Table 5.25: Audio-Audience Visual confusion matrix - 10FCV

	A	B	C	D
A	19	17	2	4
B	13	23	9	5
C	7	6	12	11
D	2	5	13	24

Table 5.26: Audio Slides confusion matrix - 10FCV

	A	B	C	D
A	12	20	8	2
B	8	32	4	6
C	2	8	18	8
D	4	4	15	21

Table 5.27: Slides-only confusion matrix - 10FCV

	A	B	C	D
A	0	29	10	3
B	0	38	11	1
C	0	22	10	4
D	0	18	13	13

Table 5.28: Emotion features-only confusion matrix - 10FCV

	A	B	C	D
A	14	16	5	7
B	6	30	9	5
C	2	11	12	11
D	1	5	13	25

Table 5.29: Emotion Speaker-Visual confusion matrix - 10FCV

	A	B	C	D
A	11	22	5	4
B	13	25	7	5
C	3	12	12	9
D	1	7	15	21

Table 5.30: Emotion Audience-Visual confusion matrix - 10FCV

	A	B	C	D
A	9	22	5	6
B	10	25	9	6
C	4	12	11	9
D	3	8	15	18

Table 5.31: Emotion Slides confusion matrix - 10FCV

	A	B	C	D
A	11	18	11	2
B	11	27	8	4
C	4	13	11	8
D	3	8	10	23

Table 5.32: All features confusion matrix - 10FCV

	A	B	C	D
A	22	16	4	0
B	10	27	9	4
C	1	15	14	6
D	1	5	13	25

We now show classification results evaluated using leave-one-out cross validation.

Table 5.33: Early-Fusion Comprehension Classification Results over 4 class distribution - LOOCV

c	Modalities	Acc	MAE	RMSE
-	**Majority Baseline**	**29.070**	**0.965**	**1.477**
1	Audio Only	**52.326**	**0.605**	**0.895**
2	Audio & Speaker Visual	**58.140**	**0.547**	**0.849**
3	Audio & Audience Visual	40.116	0.983	1.890
4	Audio & Slides	41.860	0.878	1.529
5	Slides Only	37.791	0.890	1.459
6	Emotion Features Only	42.442	0.820	1.355
7	Emotion & Speak Visual	43.605	0.779	1.233
8	Emotion & Audience Visual	39.535	0.930	1.651
9	Emotion & Slides	42.442	0.890	1.622
10	All Features	**52.907**	**0.703**	**1.207**

Table 5.33 shows 4-class classification results for all early-fusion classifiers, trained on an ordinal class classifier and evaluated using leave-one-out cross validation. Full details including the full list of features used for the training of each classifier is listed in Section 5.4.2.

Audio & Speaker features are shown to give the highest accuracy of early-fusion classification, attaining accuracy of 58.1%. Next highest accuracy for early-fusion classifiers are all features, giving an accuracy of 52.9% followed by audio-only with an accuracy of 52.3%. Confusion matrices now follow for each classifier. These give a clearer idea of the type of errors each classification model is making.

Table 5.34: Audio-only confusion matrix - LOOCV

	A	B	C	D
A	23	13	4	2
B	5	31	10	4
C	4	10	14	8
D	1	4	17	22

Table 5.35: Audio-Speaker Visual confusion matrix - LOOCV

	A	B	C	D
A	21	15	4	2
B	8	33	7	2
C	6	8	19	3
D	2	2	13	27

Table 5.36: Audio-Audience Visual confusion matrix - LOOCV

	A	B	C	D
A	21	3	14	4
B	7	7	15	7
C	11	18	20	1
D	8	10	5	21

Table 5.37: Audio Slides confusion matrix - LOOCV

	A	B	C	D
A	18	9	13	2
B	5	13	8	10
C	12	18	17	3
D	3	6	11	24

Table 5.38: Slides-only confusion matrix - LOOCV

	A	B	C	D
A	0	0	39	3
B	0	0	35	1
C	0	0	49	1
D	0	0	28	16

Table 5.39: Emotion features-only confusion matrix - LOOCV

	A	B	C	D
A	11	13	14	4
B	0	13	16	7
C	2	9	32	7
D	0	11	16	17

Table 5.40: Emotion Speaker-Visual confusion matrix - LOOCV

	A	B	C	D
A	11	18	11	2
B	0	19	8	9
C	3	14	25	8
D	0	10	14	20

Table 5.41: Emotion Audience-Visual confusion matrix - LOOCV

	A	B	C	D
A	10	14	13	5
B	4	10	11	11
C	7	8	28	7
D	1	13	10	20

Table 5.42: Emotion Slides confusion matrix - LOOCV

	A	B	C	D
A	13	11	15	3
B	3	6	19	8
C	11	7	31	1
D	6	2	13	23

Table 5.43: All features confusion matrix - LOOCV

	A	B	C	D
A	24	6	10	2
B	1	15	13	7
C	10	9	27	4
D	1	7	11	25

The confusion matrices for both 10-fold cross validation and leave-one-out cross validation evaluations give a clearer picture of the performance of each classifier than accuracy alone. From this we see that the slides-only classifier performs very poorly, with no predictions in the first comprehension class for 10-fold, and no predictions in the first two classes for leave-one-out.

Generally though, we can see from the confusion matrices we can see that the first two classes of comprehension should combine well, as should the final two classes, to convert this to a binary problem.

We now show classification results after the combination of these classes to form a binary classification problem. This classification is again performed using an ordinal class classifier and evaluated using both 10-fold cross validation and leave-one-out cross validation.

Table 5.44: Early-fusion comprehension classification results over binary distributions - 10FCV

c	Modalities	Acc	MAE	RMSE
-	**Majority Baseline**	**53.488**	**0.465**	**0.465**
1	Audio Only	**79.070**	**0.209**	**0.209**
2	Audio & Speaker Visual	**78.488**	**0.215**	**0.215**
3	Audio & Audience Visual	77.907	0.221	0.221
4	Audio & Slides	**78.488**	**0.215**	**0.215**
5	Slides Only	62.209	0.378	0.378
6	Emotion Features Only	75.000	0.250	0.250
7	Emotion & Speak Visual	76.163	0.238	0.238
8	Emotion & Audience Visual	73.837	0.262	0.262
9	Emotion & Slides	70.930	0.291	0.291
10	All Features	**78.488**	**0.215**	**0.215**

Table 5.45: Early-fusion comprehension classification results over binary distributions - LOOCV

c	Modalities	Acc	MAE	RMSE
-	**Majority Baseline**	**53.488**	**0.465**	**0.465**
1	Audio Only	**79.650**	**0.203**	**0.203**
2	Audio & Speaker Visual	**79.070**	**0.209**	**0.209**
3	Audio & Audience Visual	74.419	0.256	0.256
4	Audio & Slides	75.581	0.244	0.244
5	Slides Only	65.116	0.349	0.349
6	Emotion Features Only	76.744	0.233	0.233
7	Emotion & Speak Visual	76.163	0.238	0.238
8	Emotion & Audience Visual	75.000	0.250	0.250
9	Emotion & Slides	72.674	0.273	0.273
10	All Features	74.419	0.256	0.256

Tables 5.44 and 5.45 show early-fusion classification results over a binary range evaluated using 10-fold cross validation and leave-one-out cross validation respectively. The tables show from left-to-right the classifier number, modalities used, accuracy, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Binary classification results evaluated using 10-fold cross validation show that the best performing classifier to be the Audio-only classifier, with an accuracy of 79.07%, followed by the Audio & Speaker visual classifier and Audio & Slides classifier, both with an accuracy of 78.488%. This is also the same accuracy as the All Features classifier. For leave-one-out cross validation, the best prfomring classifier is again the Audio-only classifier, this time with an accuracy of 79.65%.

5.5.3 Late-Fusion Classification

We now investigate late fusion classification, using classifiers trained in the above Section 5.5.2. Classifiers are fused by applying weighting to the probability distributions for each class. Weightings used for each late-fused classifier are listed in results tables. Other classifiers are fused using majority vote. We use the majority class as a baseline for comparison of results. For reading of these results, numbers refer to the classifier number in the above early-fusion result tables, for example, classifiers 1, 2 & 4 refers to the audio only, audio & speaker visual and audio & slides classifiers.

Table 5.46: 7-Class Classifier - Late-Fusion

Classifiers	Accuracy	MAE	RMSE	Weighting
Majority Baseline	**29.070**	**1.157**	**2.529**	N/A
Audio-only, Audio_&_Speaker Visual & Audio_&_Slides	**52.326**	**0.680**	**1.180**	0.15-0.50-0.35
Audio_&_Speaker Visual & Audio_&_Slides	**52.907**	**0.686**	**1.233**	.50-.50

Table 5.46, listing late-fusion results for the 7-class classification problem, lists two best performing late-fusion strategies, depending on the importance of accuracy or on closeness of predictions to the actual. For the first displayed late-fusion classifier, classifiers 1, 2 and 4 are fused using weighted majority vote, applying weightings of 0.5 for audio & speaker movement features, 0.35 for audio & slides and 0.15 for audio-only. The other classifier in this table is fused comprising of the audio and speaker movement classifier and the audio and slides classifier, using an equal, 50-50 weighting of probability distributions. These weightings were arrived at by running a series of experiments using different weightings and selecting the best performing weights.

Table 5.47: 4-Class Classifier - Late-Fusion

Modalities	Accuracy	MAE	RMSE	Weighting
Majority Baseline	**29.070**	**0.965**	**1.477**	N/A
Audio-only, Audio_&_Speaker Visual & Audio_&_Slides	**58.140**	**0.535**	**0.779**	.40-.35-.25
Audio-only, Audio_&_Speaker Visual & Audio_&_Audience Visual	55.223	0.616	1.012	.40-.35-.25
Audio-only, Audio_&_Speaker Visual & Audio_&_Audience Visual & Audio_&_Slides	56.395	0.558	0.826	.30-.275-.225-.20

Table 5.47 shows late-fusion results for the four-class classification range. We obtain best performance using the early-fusion audio-only, audio & speaker movement and audio & slides classifiers. These classification outputs are fused using weighted majority vote probability distributions from each classification output. Weightings for each classifier are shown in Table 5.47. The next best performing late-fusion classifier for four-class classification is fused on early-fusion classification outputs from audio-only, audio & speaker movement, audio & audience movement and audio & slides. These early-fused classifiers are again fused using weighted vote of probability distributions. These weightings were arrived at by running a series of experiments using different weightings and selecting the best performing weights.

Table 5.48: Binary Classifier - Late-Fusion

Modalities	Accuracy	MAE	RMSE	Weighting
Majority Baseline	**53.488**	**0.465**	**0.465**	N/A
Audio-only, Audio_&_Speaker Visual & Audio_&_Slides	**80.814**	**0.192**	**0.192**	Simple Majority Vote
Audio_&_Audience Visual & Audio_&_Slides & Emotion&Speaker Visual	80.233	0.198	0.198	Simple Majority Vote
Audio-only & Audio_&_Speaker Visual	79.651	0.203	0.203	Simple Majority Vote
Audio_&_Speaker Visual & Audio_&_Slides	76.744	0.233	0.233	Simple Majority Vote
Audio_&_Audience Visual & Audio_&_Slides	80.233	0.198	0.198	Simple Majority Vote

Table 5.48, late-fusion of the binary classification problem, shows the best performing classifiers on this problem. The best performing classifier is late fused on audio-only, audio & speaker movement, and audio & slides. This uses a simple majority vote approach from each early-fusion classifier, in which the class for which most early-fusion classifiers output to is taken as the final result. For instances of a tie, the class with the highest probability distribution is taken as the final resulting class.

Following this, Figures 5.4 and 5.5 plot classifications from the best performing late-fusion classifiers listed above for classification over the 7-class and 4-class ranges.

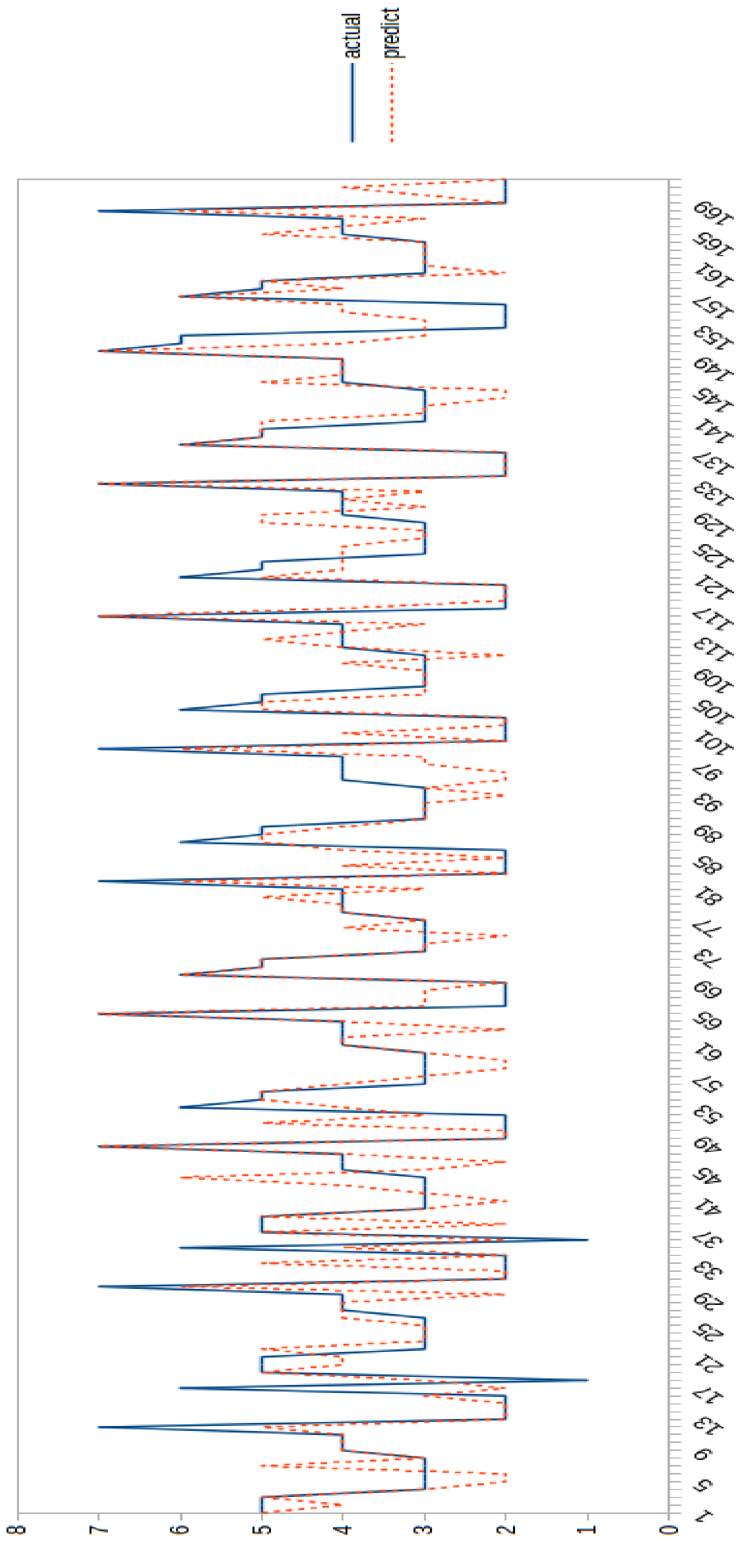


Figure 5.4: Plotted classifications over the full 7 class distribution. These are the results of late fusion of classifiers. The numbers along the x-axis are the video segment numbers, time-plotted over the full data-set. The numbers along the y-axis are the comprehension levels, ranging from 1 to 7. As explained by the legend, the solid blue line is the actual comprehension level for that segment while the broken red line is the prediction for that segment.

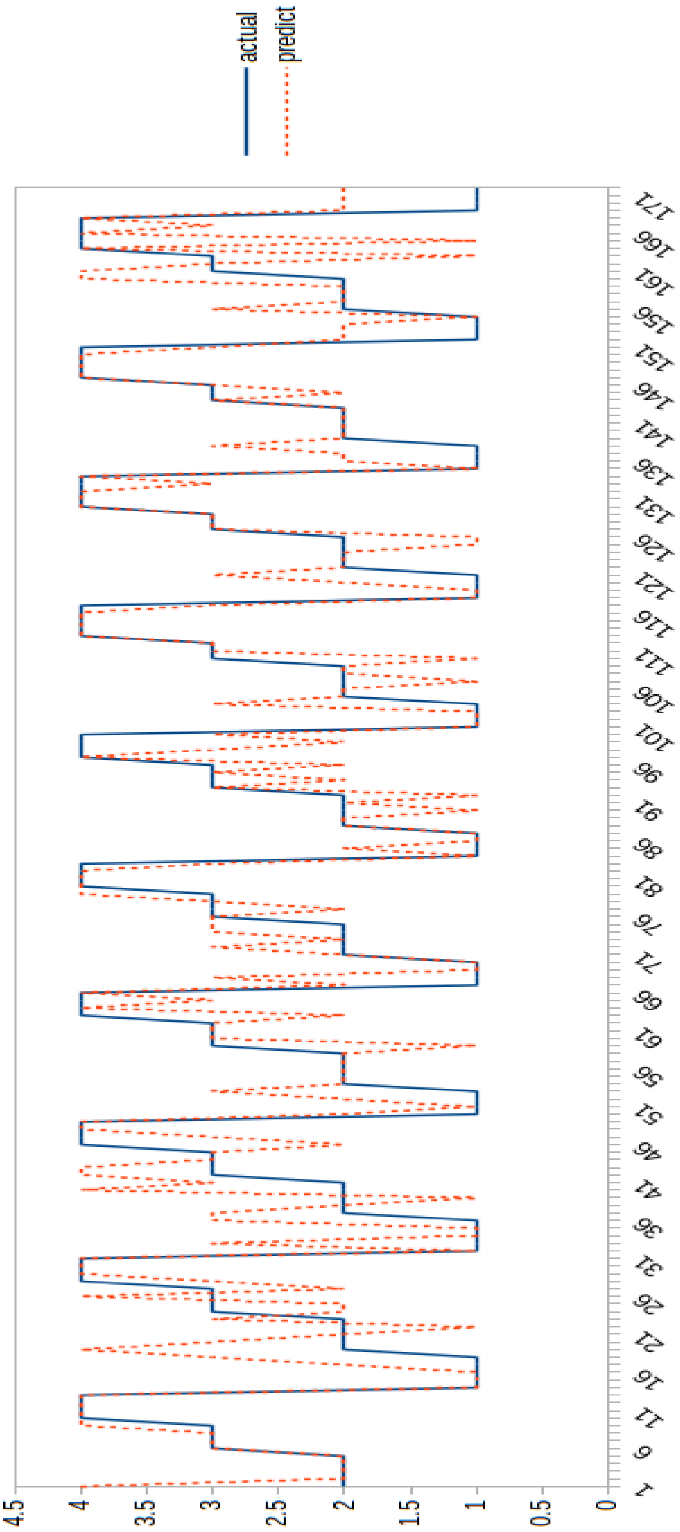


Figure 5.5: Plotted classifications over the combined 4 class distribution. These are the results of late fusion of classifiers. The numbers along the x-axis are the video segment numbers, time-plotted over the full data-set. The numbers along the y-axis are the comprehension levels, ranging from 1 to 4. As explained by the legend, the solid blue line is the actual comprehension level for that segment while the broken red line is the prediction for that segment.

5.5.4 Alternative Classification for Binary Problem

We now perform alternative classification over the binary classification problem using a Rotation Forest classification algorithm. This algorithm has shown to give good results on similar binary classification tasks, hence we use this to classify this concept over a binary problem giving that we are no longer constrained by an ordinal problem and binary classification is easier than multi-class classification. We again evaluate these using both 10-fold cross validation and leave-one-out cross validation.

Table 5.49: Binary Classification - Rotation Forest - 10FCV

c	Modalities	Acc	MAE	RMSE
-	**Majority Baseline**	**53.488**	**0.465**	**0.465**
1	Audio-Only	83.140	0.169	0.169
2	Audio & Speaker Visual	83.140	0.169	0.169
3	Audio & Audience Visual	83.140	0.169	0.169
4	Audio & Slides	81.395	0.186	0.186
5	Slides Only	66.279	0.337	0.337
6	Emotion Features Only	78.488	0.215	0.215
7	Emotion & Speaker Visual	77.326	0.227	0.227
8	Emotion & Audience Visual	77.907	0.221	0.221
9	Emotion & Slides	76.744	0.233	0.223
10	All Features	**85.465**	**0.145**	**0.145**

Table 5.50: Binary Classification - Rotation Forest - LOOCV

c	Modalities	Acc	MAE	RMSE
-	**Majority Baseline**	**53.488**	**0.465**	**0.465**
1	Audio-Only	78.488	0.215	0.215
2	Audio & Speaker Visual	81.395	0.186	0.186
3	Audio & Audience Visual	75.000	0.250	0.250
4	Audio & Slides	83.140	0.169	0.169
5	Slides Only	62.791	0.372	0.372
6	Emotion Features Only	75.581	0.244	0.244
7	Emotion & Speaker Visual	80.233	0.198	0.198
8	Emotion & Audience Visual	76.744	0.233	0.233
9	Emotion & Slides	77.907	0.221	0.221
10	All Features	**81.977**	**0.180**	**0.180**

Rotation Forest (Rodriguez et al., 2006) is an algorithm for generating classifier ensembles based on feature extraction. To create the training data for a base classifier, the feature set is randomly split into K subsets (K is a parameter of the algorithm) and Principal Component Analysis (PCA) is applied to each subset. All principal components are retained in order to preserve the variability information in the data. Thus, K axis rotations take place to form the new features for a base classifier. The idea of the rotation approach is to encourage individual accuracy and diversity simultaneously within the ensemble. Diversity is promoted through the feature extraction for each base classifier. Decision trees were chosen here by the algorithm because they are sensitive to rotation of the feature axes, hence the name "forest." Accuracy is sought by keeping all principal components and also using the whole data set to train each base classifier.

Tables 5.49 and 5.50 show results for the binary classification task using a Rotation Forest classification algorithm, evaluated using 10-fold cross validation (10FCV) and Leave-One-Out cross validation (LOOCV). This takes advantage of the fact that we are no longer using an ordinal scale for classification, and thus are no longer limited to an ordinal classifier. These results show that the use of this classification algorithm gives consistently greater results over the binary task than the Ordinal Class Classifier used for multi-class classification. Table 5.49 shows classification results over all features and yields a maximum result of 85.465%. This is a very good result which proves that comprehension is a concept which can be accurately classified during audio-visual presentations.

5.5.5 Correlation of Individual Features to Comprehension

In this section we calculate the linear correlations between individual audio-visual features extracted from the dataset and final human annotations on comprehension. This is calculated by aligning values for each of these features with the final annotation for comprehension for each video segment. This was calculated using the Pearson's Correlation Coefficient in order to discover the linear correlation.

Tables 5.51 and 5.52 above shows linear correlations between extracted audio-visual features and average audience comprehension levels for each video segment, taken from the human annotated ground-truth. As can be seen from these results,

Table 5.51: Multimodal Features - Comprehension : Linear Correlation

Multimodal Feature	$r =$	**Multimodal Feature**	$r =$
Pitch Max	0.0867	F3 - F4	-0.0827
Pitch Range	0.0715	F4 / F1	**-0.3576**
Pitch Mean	0.0436	Pauses Per Minute	-0.0257
Pitch Variance	0.011	Slide Clutter Range	**0.2329**
Pitch Std Dev	0.0299	Slide Clutter Mean	**0.2928**
Intensity Max	0.0877	Slide Clutter Variance	**0.2256**
Intensity Range	0.1203	Slide Clutter Std Dev	**0.2079**
Intensity Mean	0.0642	Avg Words per Slide	-0.0477
Intensity Variance	0.014	Avg Word Length	0.127
Intensity Std Dev	0.0057	Avg Lines per Slide	-0.1015
Speech Rate	-0.0453	Face Movement Range	**0.2349**
Articulation Rate	**0.2464**	Face Movement Mean	**0.39**
ASD	**-0.2394**	Face Movement Variance	**0.3743**
F1	0.0879	Face Movement Std Dev	**0.3494**
F2 - F1	-0.1848	Speaker Motion Range	-0.026

Table 5.52: Multimodal Features - Comprehension : Linear Correlation

Multimodal Feature	$r =$	**Multimodal Feature**	$r =$
Speaker Motion Mean	-0.0179	pcm_zcr_sma_de_linregerrQ	0.0958
Speaker Motion Variance	-0.0576	pcm_zcr_sma_de_stddev	0.1061
Speaker Motion Std Dev	-0.0291	mfcc_sma[10]_min	0.1458
Aud Face Counts Range	-0.0677	mfcc_sma_de[7]_linregc2	**0.2655**
Aud Face Counts Variance	0.0252	mfcc_sma[12]_stddev	-0.1956
Aud Face Counts Std Dev	0.0266	mfcc_sma[2]_min	**-0.2957**
Audience Motion Range	-0.0542	mfcc_sma[10]_skewness	**0.2043**
Audience Motion Mean	-0.0667	mfcc_sma_de[2]_stddev	**0.2316**
Audience Motion Variance	-0.069	mfcc_sma_de[2]_linregerrQ	**0.2271**
Audience Motion Std Dev	-0.0778	mfcc_sma[2]_amean	**-0.2817**
mfcc_sma_[11]_linregc2	**-0.2329**	mfcc_sma_de[12]_stddev	-0.1173
mfcc_sma[3]_skewness	-0.1331	mfcc_sma_de[12]_linregerrQ	-0.126
mfcc_sma[11]_amean	**-0.2203**	mfcc_sma[3]_min	**-0.2234**
mfcc_sma[12]_skewness	0.0917	F0_sma_de_max	-0.0925
mfcc_sma_de[3]_skewness	-0.0647	mfcc_sma[7]_amean	**0.2703**

articulation rate, clutter of slides and facial movement of the speaker can all be said to affect audience comprehension to some degree.

5.5.6 Relation to Speaker Ratings and Audience Engagement

We next test for any potential correlations between the concepts of audience comprehension and engagement. The Pearson's linear correlation measures the correlation coefficient between human-annotated levels of audience engagement and comprehension levels. This is calculated by aligning human annotations for audience comprehension and for audience engagement for each video segment, and then calculating the Pearson's Correlation Coefficient. The linear correlation of 0.0487 shows that there is no real correlation between these two concepts, confirming that it is possible for an audience to be engaged with a presentation while not fully comprehending its material.

Further correlations between previously annotated speaker ratings and audience comprehension levels, again by aligning annotations for each video segment and calculating the Pearson's Correlation Coefficient, show a linear correlation of 0.1213. This is a very weak positive correlation, indicating that what people typically perceive to be aspects of good public speaking do not always correlate well to actual levels of comprehension among the audience to such speech. This means that it is possible for a speaker to be using presentation techniques considered most engaging for the audience, whilst simultaneously failing to make the material comprehensible for the audience.

5.6 Summary

In this chapter we explored the ability to predict the potential for a presentation to be comprehended by the audience. Audio-visual features were extracted from the presenter and the audience to academic presentations. Visual features were also extracted from the slides for each presentation. We experimented with the use of early-fusion and late-fusion techniques for the training of a classifier for this task. Separate classifiers were trained on each set of features and later fused using a late-fusion strategy. Classifiers were also trained over all feature sets to discover the best fusion strategies for tasks of this nature. Little difference was found between the results of early-fusion and late-fusion techniques.

Overall late-fusion strategies performed marginally better on classification tasks over the 7-class distribution. Over a 4-class distribution, late-fusion strategies again outperformed early-fusion strategies. However, over a binary class distribution, early-fusion techniques performed slightly better when trained over the whole set of features. Overall we have achieved good results on these classification tasks.

We have also investigated correlations between speaker ratings / audience engagement levels and audience comprehension levels. We discovered no real correlation, confirming that it is possible for an audience to be fully engaged with a presentation whilst not fully comprehending the material.

Having successfully classified areas of good speaker ratings, high audience engagement, instances of intentional or unintentional emphasised speech and the potential of a speaker to be comprehended by an audience, we now explore the potential of these high-level paralinguistic features to be used to improve summarisation of academic presentations. In the next chapter, we develop an algorithm to automatically summarise presentations using these features in addition to linguistic features such as keyword information. We then perform a number of experiments designed to evaluate the effectiveness of this summarisation approach.

Chapter 6

Generation and Evaluation of Presentation Summaries

In this chapter we describe the process of automatically generating summaries of academic presentations. Summaries are created and evaluated based on both standard linguistic features and also the paralinguistic features introduced in the preceding chapters. The algorithms used for creating summaries are explained, including details of the individual factors used to create them. We also introduce techniques for the evaluation of generated summaries. Finally, we present results and analysis of our evaluation of the summaries and draw conclusions.

6.1 Introduction

Presentations from academic conferences are potentially of great interest to researchers working in the area of the presentations. However, viewing full academic presentations can be time consuming, and, while keywords can give an indication of just how relevant each presentation is to the researcher, they do not know the value of a presentation until they actually watch it. There are other times where they may wish to catch up on an academic presentation, and simply do not have the time to spare to watch it in its entirety.

Suitably constructed automatically generated summaries of academic presentations from international conferences could save users from needing to view full presentations in order to gauge their utility and access the information contained in them.

In order to automatically generate presentation summaries of the type suggested here, what features should be used as the basis of such summaries? Content found to be most engaging, emphasised and comprehensible for an audience, as described earlier in this thesis, seems like a good place to start, as there seems little point in summarising presentation material if we do not capture these most interesting and comprehensible parts of the presentation.

However, summaries would also need to capture the important parts, and information relevant to the user's interest in the presentation. For this we would want to ensure that such summaries capture the important keywords from each presentation, and also to ensure that summaries take into account all parts of a presentation. For example, if the most engaging parts of a presentation are all to be found at the beginning of the talk, engaging summaries built from this would be close to useless is they just contain all of the most engaging parts of that talk whilst missing out on less engaging, yet critical aspects from the end of the talk such as the results and conclusions found.

In this chapter we address our fifth and final Research Question:

5. **'Can areas of special emphasis provided by the speaker, combined with detected areas of high audience engagement and high levels of audience comprehension, be used as a component in the effective summarisation of academic presentations?'**

To reliably answer this research question we introduce a mechanism for creation of summaries of academic presentations, and perform a comprehensive evaluation of these summaries. In this regard, we make use of eye-tracking as participants watch full presentations and watch separate presentation summaries. Eye-tracking is suitable for evaluating video summaries and is performed for this evaluation because, as shown in previous work, an increased number of shorter fixations is consistent with higher cognitive activity (attention), while a reduced number of longer fixations is consistent with lower attention (Rayner and Sereno, 1994). This shows that more engaging videos and video summaries would be expected to induce a higher number of shorter fixations from participants than non-engaging presentations and summaries. This enables us to study whether generated summaries have the desired (more engaging) effect on levels of attention / engagement of participants as they watch presentation summaries.

For further evaluation, presentation summaries are also compared with the use of an enhanced digital video browser, previously found to be very effective for assisting users in gaining a quick overview of presentations and lectures (Li et al., 2000).

This chapter is structured as follows: The first section introduces our method for the automatic generation of presentation summaries. Using features developed in earlier chapters, we seek to make our summaries engaging, emphasised and comprehensible, yet also to contain all important and relevant parts of the presentation. To ensure this, summaries are built by rating all parts of a presentation by importance of their inclusion in a summary. Full presentations are broken into sentences, of which each is rated based on its classified ratings for engagement, emphasis, comprehension, and also the number of keywords contained in the sentence.

Presentation summaries are only as effective as they are found to be by their audience. In the next part of this chapter we provide a detailed evaluation of our automatically generated presentation summaries in order to examine their effectiveness using eye-tracking methods as outlined above, and in comparison to a video browsing application. Additionally, questionnaires are used to elicit user feedback on these summaries.

6.2 Creation of Presentation Summaries

This section describes the steps involved in our generation of presentation summaries. Presentations had been processed by (Spoken Data Video Processing), who extracted ASR transcripts and significant keywords associated with these transcripts. Summaries were generated using these ASR transcripts, significant keywords extracted from these transcripts, and annotated values for 'good' public speaking techniques, audience engagement, intentional or unintentional speaker emphasis and the speakers potential to be comprehended. The summarisation process can also be visualised as shown in Figure 6.1.

1. ASR outputs are segmented using the pause information in the transcripts, which indicate start and end times for each spoken phrase. These segments provide a basis for the segmentation of presentations at the phrase-level, with significant phrases selected for inclusion in the summary.

2. We first apply a ranking for each phrase based on the number of keywords, or words of significance, contained within it. For the first set of baseline summaries, we generate summaries by using the highest ranking phrases.

3. Speaker Rating's are halved before applying this ranking to each phrase. We half the values for speaker ratings so as not to overvalue this feature, as these values are already encompassed for classification of audience engagement levels, see Section 4.1.4.2.

4. Following this, audience engagement annotations are also applied to phrases. We take the final annotated engagement level and apply this value to each phrase contained within each segment throughout the presentation.

5. As emphasis was not annotated for all videos, we use automatic classifications for intentional or unintentional speaker emphasis. For each classification of emphasis, we apply an additional value of 1 to the phrase containing that emphasised part of speech.

6. Finally, we use the human annotated values for audience comprehension throughout the dataset. Once again the final comprehension value for each segment is also applied to each phrase within that segment. For weightings of paralinguistic feature values, we choose to half the Speaker Rating annotation, while choosing to keep the original for the other annotations for engagement, emphasis and comprehension. This is because Speaker Ratings are already used for classification of engagement. Points of emphasis receive a value of 1, while keywords receive a value of two, in order to give importance to the role of keywords in the summary generation process.

   To generate the final set of video summaries, the highest scoring phrases in the set are selected. To achieve this, the final ranking for each phrase is normalised to between 0 and 1.

7. By then assigning an initial threshold value of 0.9, and reducing this by 0.03 on each iteration, we select each sentence with a ranking above that threshold value. By calculating the length of each selected sentence, we can then apply a minimum size to our generated video summaries. Using this method we can choose a minimum summary length, allowing us to avoid summaries which are too detailed or not sufficiently detailed. Final selected segments are then joined together to generate small, medium and large summaries for each presentation. The temporal order of segments within summaries is preserved.

Full details of final summary lengths are shown in Table 6.1 and Table 6.2. The first column lists the name of the presentation, second column the overall length of the full presentation, third column the length of the generated short summary. The fourth column is the length of the medium summary and the fifth column is the length of the long summary. The remaining three columns are the lengths of the summary as a percentage of the full presentation.

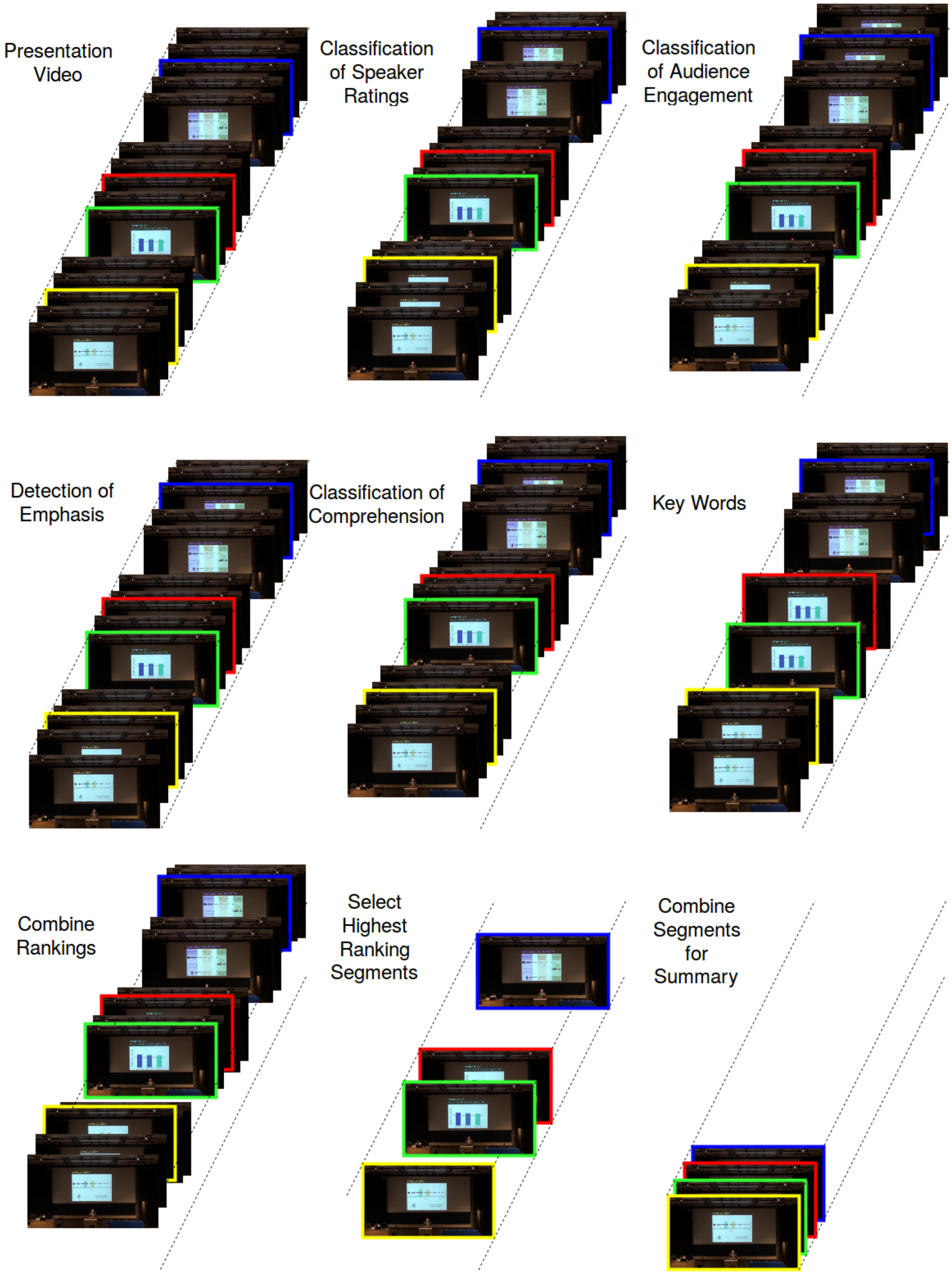

Presentation Video
Classification of Speaker Ratings
Classification of Audience Engagement
Detection of Emphasis
Classification of Comprehension
Key Words
Combine Rankings
Select Highest Ranking Segments
Combine Segments for Summary

Figure 6.1: Visualisation of Video Summarisation Process

Algorithm 2 Generate Summaries

```
for all $_1Sentence \rightarrow S$ do
    if $S\_contains\_Keyword$ then
        $_2S \leftarrow S + 2$
    end if
    $Engagement \rightarrow E$
    $SpeakerRating \rightarrow SR$
    $Emphasis \rightarrow Es$
    $Comprehension \rightarrow C$
    $_3S \leftarrow S + E$
    $_4S \leftarrow S + SR/2$
    $_5S \leftarrow S + Es$
    $_6S \leftarrow S + C$
end for
while $Summary < length$ do
    if $S \geq Threshold$ then
        $_7Summary \leftarrow S$
    end if
end while
```

Table 6.1: Video Summary Lengths 1

Video	Length	Short	Medium	Long	Short_%	Medium_%	Long_%
intonation1	1351	326	471	570	24.1%	34.9%	42.2%
intonation2	1030	214	335	414	20.8%	32.5%	40.2%
intonation4	1188	221	403	532	18.6%	33.9%	44.8%
intonation5	1038	211	362	400	20.3%	34.9%	38.5%
intonation6	891	164	287	348	18.4%	32.2%	39.1%
plen1	1166	221	387	449	19%	33.2%	38.5%
plen2	966	207	278	371	21.4%	28.8%	38.4%
plen3	1030	206	401	401	20%	38.9%	38.9%
plen4	996	204	309	379	20.5%	31%	39.2%
plen5	816	170	275	339	20.8%	33.7%	41.5%
plen6	776	166	224	302	21.4%	28.9%	38.9%
plen10	1048	232	321	412	22.1%	30.6%	39.3%
plen11	956	204	352	406	21.3%	36.8%	42.5%

Table 6.2: Video Summary Lengths 2

Video	Length	Short	Medium	Long	Short_%	Medium_%	Long_%
plen12	1100	276	349	500	25.1%	31.7%	45.5%
prp1	873	212	254	341	24.3%	29.1%	39.1%
prp2	1167	294	395	473	25.2%	33.8%	40.5%
prp3	1219	247	404	530	20.3%	33.1%	43.5%
prp4	1004	222	283	371	22.1%	28.2%	37%
prp5	767	140	219	328	18.3%	28.6%	42.8%
prp6	1190	256	405	473	21.5%	34%	39.7%
slavicp1	1301	264	489	489	20.3%	37.6%	37.6%
slavicp2	995	209	350	383	21%	35.2%	38.5%
slavicp3	1071	272	307	437	25.4%	28.7%	40.8%
slavicp4	1075	228	310	427	21.2%	28.8%	39.7%
slavicp5	1040	226	325	575	21.7%	31.3%	55.3%
speechRT1	809	145	230	338	17.9%	28.4%	41.8%
speechRT2	862	186	245	350	21.6%	28.4%	40.6%
speechRT3	928	224	276	439	24.1%	29.7%	47.3%
speechRT4	882	172	259	334	19.5%	29.4%	37.9%
speechRT5	794	149	213	301	18.8%	26.8%	37.9%
speechRT6	936	182	281	390	19.4%	30%	41.7%

Table 6.3: Final Video Summary Percentages

Summary Size	Summed Size %	Total Summaries	Average Size %
Small	656.4	31	21.2
Medium	983.1	31	31.7
Large	1269.2	31	40.9

6.3 Evaluation of Video Summaries

Initial evaluations of our automatically generated presentation summaries were performed by performing eye-tracking of participants as they watched a full presentation video and an automatically generated video summary. A complimentary comparison was also performed examining the use of an enhanced digital video browser tool, which removed pauses and allowed playback to be increased to 250% of the norm to support manual skimming, The video browser application was previously found to be very effective for gaining a quick overview of conference presentations (Li et al., 2000). For additional evaluations, we crowdsourced human evaluation tasks by asking workers to watch summaries generated using the keywords taken from the ASR transcripts alone and summaries generated using all available information.

The question being addressed by the eye-tracking study, was whether or not participants spend a larger proportion of the time viewing slides during summaries than full presentations? Previous work has shown that an increased number of shorter fixations is consistent with higher cognitive activity (attention), while a reduced number of longer fixations is consistent with lower attention (Rayner and Sereno, 1994). This allows us to understand clearly whether generated summaries have any effect on levels of attention / engagement of participants as they watch presentation summaries.

Summaries for eye-tracking evaluations were built using the human annotations of the videos in order to assume best-case scenario for automatic classifications. We also build summaries using automatic classification results in addition to keywords, automatic classifications with no keywords, automatic classifications using audio-only features in addition to keywords, and automatic classifications using visual-only features in addition to keywords for questionnaire evaluations to evaluate the importance of audio and visual features for generation of automatic presentation summaries.

The overarching experiment plan includes eye-tracking of participants as they watch full presentations and summaries followed by answering of a questionnaire on each of these. Heat Maps and Gaze Plots were created from the eye-tracking results. As described above, summaries were built using subsets of available features and compared through the use of questionnaires. Summaries were also compared with an enhanced digital video browser for effectiveness and ease of use for gaining a quick, clear and concise overview of a presentation in a short time frame.

Four test videos were used for evaluation. These test videos were chosen based on earlier human annotations for speaker ratings, audience engagement, and audience comprehension. We used the videos found to be most engaging, most comprehensible, best speaker ratings, and least engaging.

6.3.1 Gaze-Detection: Instructions to Participants

Instructions given to participants in gaze-detection evaluations were simple and limited to what was necessary. After a short introduction on how the gaze-detection equipment works, participants were instructed to avoid bringing their hands up to their faces and covering their eyes as this would affect the detections of gaze. Participants in these evaluations were also instructed that the researcher would be sitting at a desk behind them to their right, but would not be watching them. The only other instruction participants received was to watch the presentations and answer the questionnaire that followed each. Given participants further instructions, such as where to focus or not to focus, was avoided for fear of influencing results. The purpose of these evaluations was to have participants watch presentations in as natural a way as possible so that their gaze could be evaluated for engagement.

6.3.2 Gaze-Detection Evaluation

Eye-tracking was performed for evaluation of presentation summaries. Participants watched one full conference presentation whilst having their eye-movements tracked and then answered a questionnaire on the same. Participants also watched a separate presentation summary, again whilst having their eye-movements tracked and again answered a number of questions on same. For eye-tracking experiments, eight different condition tests were developed, with a minimum of 4 participants per test. These were based on the four presentations and summaries used for evaluation and by mixing the order of full presentation of summary to be viewed first in order to reduce biases and/or factors of fatigue from influencing results.

Participants who began by watching a full presentation video concluded by watching a summary of a different presentation, while participants who began by watching a summary of a presentation concluded by watching a different full presentation. This was again to reduce biases and factors of fatigue from influencing results and to ensure that eye-tracking results were not influenced by a participant already having prior knowledge of a presentation.

In the below tables we use the terms video, version and scene. Video is self explanatory, version refers to either the full version or the summary version, and scene refers to the scenes into which the videos were separated for analysing results. These results separated videos into the slides scene and the speaker scene, to compare values for fixations over the presentation slides and over the presenter themselves. The final three columns for results on each video are the Mean Difference, the Standard Error, and the Statistical Significance. Differences are regarded as being statistically significant if they show a Statistical Significance value of 0.05 or less.

Table 6.4 shows the core values for eye-tracking measurements per video, version and scene, from which the above version tables for videos 1 - 4 are calculated. For example, 3.2.2 in the V.V.S. field indicates Video 3, Version 2, Scene 2. Measurements obtained include number of Fixations, Mean length of fixations, Total Sum of fixation lengths, percentage of time fixated per scene, average number of fixation counts and Number of fixation counts per 100 seconds.

The following Tables 6.5, 6.6, 6.7, and 6.8 each show results for each of the selected test videos:

Table 6.5 shows a statistically significant ($p < 0.05$) difference between the

Table 6.4: Totals per video, version, scene (V.V.S = Video,Version,Scene)

V.V.S.	FD.N	FD.M	FD.S	%	FC.S	FCp100
1.1.1	354.875	0.421	143.874	66.608	354.875	164.294
1.1.2	62.500	0.959	55.296	25.600	62.500	28.935
1.2.1	1223.500	0.446	536.485	55.537	1223.500	126.978
1.2.2	299.500	1.246	329.468	34.106	299.500	31.004
2.1.1	248.125	0.590	145.479	66.127	248.125	112.784
2.1.2	45.125	1.244	6.518	25.690	45.125	20.511
2.2.1	923.125	0.666	601.053	68.849	923.125	105.742
2.2.2	165.500	1.077	160.774	18.416	165.500	18.958
3.1.1	140.125	0.429	60.123	40.623	140.125	94.679
3.1.2	53.750	1.337	69.969	47.276	53.750	36.318
3.2.1	596.500	0.580	332.088	43.297	596.500	77.771
3.2.2	294.875	1.329	324.413	42.296	294.875	38.445
4.1.1	345.125	0.447	149.583	79.144	345.125	182.606
4.1.2	40.375	0.611	24.523	12.975	40.375	21.362
4.2.1	1194.750	0.505	594.019	63.464	1194.750	127.644
4.2.2	219.625	0.814	181.351	19.375	219.625	23.464

Table 6.5: Eye-tracking Video 1 - slides scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.02	0.05	0.655
summ	**full**	**percent**	**Scheffe**	**11.07**	**4.97**	**0.043**
summ	**full**	**FCp100**	**Scheffe**	**37.32**	**14.32**	**0.021**

summary and full versions, for the percentage of time fixated per scene and the number of fixations per 100 seconds. As also demonstrated in the core figures in table 6.4, this indicates a significant difference in the amount of time users spent fixating on the slides while watching the summary than the full video. This indicates that users found there to be a much higher concentration of new information during the summary than during full presentations.

Table 6.6: Eye-tracking Video 2 - slides compared scene by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.08	0.08	0.337
summ	full	percent	Scheffe	-2.72	7.90	0.735
summ	full	FCp100	Scheffe	7.04	11.16	0.538

Again in Table 6.6, key differences are observed in the average fixation duration per scene, and to a lesser extent in the fixation count per 100 seconds, more clearly visible from the core figures available in Table 6.4, neither of these differences are statistically significant however, meaning there is not too much difference between engagement and focus levels for full presentation and for summaries of video 2, prp1.

Table 6.7: Eye-tracking Video 3 - slides scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	**full**	**FD.M**	**Scheffe**	**-0.15**	**0.05**	**0.009**
summ	full	percent	Scheffe	-2.67	6.54	0.689
summ	full	FCper100	Scheffe	16.91	13.54	0.232

Table 6.7 shows that there is a significant difference ($p < 0.05$) in the mean fixation length over the slides during the summary than during the full presentation. As can be seen by looking at the core figures in Table 6.4, the averaged fixation duration is much shorter for fixations on slides. This indicates that users tend to read over slides in order to follow and absorb printed information. Shorter fixations is also consistent with increased cognitive activity.

Table 6.8: Eye-tracking Video 4 - slides scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.06	0.05	0.266
summ	**full**	**percent**	**Scheffe**	**15.68**	**6.42**	**0.028**
summ	**full**	**FCp100**	**Scheffe**	**54.96**	**16.98**	**0.006**

Table 6.8 shows a statistically significant ($p < 0.05$) difference between the summary and full versions, for the percentage of time fixated per scene and the number of fixations per 100 seconds. As demonstrated in the core figures in Table 6.4, this indicates a significant difference in the amount of time users spent fixating on the slides while watching the summary than the full video. This indicates that users found there to be a much higher concentration of new information during the summary than the full version. The is also a big, though not quite statistically significant difference in mean fixation duration for video 4, indicating that participants found this summary video to be more engaging than the full presentation.

Full results from these experiments can be found in Appendix B.1.

6.3.3 Further Evaluations of Gaze

By combining the slides and speaker scene's from the last set of eye-tracking evaluations into one combined attention scene, and comparing with the whole scene, we can evaluate further differences between full presentations and summaries in how much they both entice full attention to the presentations in progress. In these results, our scene refers to either the combined attention scene (slides and presenter) or the overall full scene.

Table 6.9: Totals per video, version, scene (V.V.S = Video,Version,Scene)

V.V.S.	FD.N	FD.M	FD.S	%	FC.S	FCp100
1.1.1	432.500	0.492	203.986	94.438	432.625	200.289
1.1.2	417.370	0.495	199.170	92.208	417.375	193.229
1.2.1	1580.000	0.582	889.486	92.079	1580.000	163.561
1.2.2	1523.000	0.587	865.952	89.643	1523.000	157.66
2.1.1	311.870	0.695	209.284	95.129	311.875	141.761
2.1.2	293.250	0.710	201.996	91.816	293.250	133.295
2.2.1	1153.120	0.709	780.864	89.446	1153.125	132.087
2.2.2	1091.120	0.724	761.826	87.265	1091.125	124.987
3.1.1	224.370	0.620	135.407	91.491	224.375	151.605
3.1.2	193.870	0.694	130.091	87.899	193.875	130.997
3.2.1	1076.750	0.641	643.995	83.963	1076.750	140.385
3.2.2	891.370	0.800	656.500	85.593	891.375	116.216
4.1.1	406.370	0.431	169.388	89.624	406.375	215.013
4.1.2	385.500	0.466	174.105	92.119	385.500	203.968
4.2.1	1591.250	0.536	832.177	88.908	1591.25	170.005
4.2.2	1414.370	0.561	775.370	82.839	1414.375	151.108

Table 6.9 shows the averaged core values for eye-tracking measurements per video, version and scene. Videos are listed 1 to 4, with plen2 as video 1, prp1 as video 2, prp5 as video 3, and speechRT6 as video 4. Version is listed 1 to 2, with the video summary as version 1, and the full video as version 2. Scene is also listed 1 to 2, with the overall scene as scene 1 and the attention scene, the area around the slides and the speaker, as scene 2. Measurements obtained include number of Fixations, Mean length of fixations, Total Sum of fixation lengths, percentage of time fixated per scene, average fixation counts and Number of fixations per 100 seconds.

Also, we can see that participants for all videos had a higher number of fixations for the overall scene than the attention scene, as the overall scene encompasses the attention scene, yet had a much lower mean fixation duration. This indicates that participants tended to make many, short fixations on this scene, indicating that participants tended to be more engaged for fixations in this area.

Again, from Table 6.9, we can see that participants consistently spend a higher proportion of the time fixating on the scene for summaries than for the full presentation video. This is repeated to an even larger extent for Fixation Counts per 100 seconds, where this figure is consistently higher for summaries than for full presentations. Again, this is evidence of increased levels of audience engagement for video summaries than for full presentation videos.

We can see that the number of fixations per second is consistently higher for video summaries while the mean fixation length is consistently shorter for summaries. As previous work has shown that an increased number of shorter fixations is consistent with higher cognitive activity (attention), while a reduced number of longer fixations is consistent with lower attention (Rayner and Sereno, 1994), this shows that all video summaries attract higher attention levels for summaries than for full presentations.

Table 6.10: Eye-tracking Video 1 - scene by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.09	0.06	0.163
summ	full	percent	Scheffe	2.36	1.68	0.181
summ	**full**	**FCp100**	**Scheffe**	**36.73**	**17.12**	**0.050**

Table 6.10 shows a statistically significant difference between the summary and full versions, for the number of fixations per 100 seconds. Taking these results in conjunction with the overall core figures in Table 6.9 show that video 1 summary is more engaging than the overall presentation video for video 1.

Again in Table 6.11, key differences are observed in the average fixation duration per scene, and to a lesser extent in the fixation count per 100 seconds, more

Table 6.11: Eye-tracking Video 2 - scene by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.01	0.08	0.865
summ	**full**	**percent**	**Scheffe**	**5.68**	**2.41**	**0.033**
summ	full	FCp100	Scheffe	7.04	14.51	0.516

clearly visible from the figures in Table 6.9, neither of these differences are statistically significant. Participants spend a statistically significant higher proportion of their time fixating on the attention scene for video summaries than for full video presentations.

Table 6.12: Eye-tracking Video 3 - scene by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.02	0.09	0.813
summ	full	percent	Scheffe	7.53	3.63	0.057
summ	full	FCper100	Scheffe	11.22	15.24	0.474

Table 6.12 shows that there is no statistically significant difference between the two scene's of the video, however, there is a large, but not significant difference in the percentage of time spent fixating on the attention scene during the video summary compared with during the full video presentation.

Table 6.13: Eye-tracking Video 4 - scene by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	Scheffe	-0.11	0.06	0.080
summ	full	percent	Scheffe	0.72	3.62	0.846
summ	**full**	**FCp100**	**Scheffe**	**45.01**	**15.27**	**0.011**

Table 6.13 shows a statistically significant ($p < 0.05$) difference between the summary and full versions for the number of fixations per 100 seconds. There is also a big difference between version for the mean fixation duration. As engagement and focus has been found to be consistent with a high number of short fixations, this indicates that users found the summary of this video to be much more engaging than the full presentation. Further, this video had previously been found to be least engaging by our classifiers before summarisation, indicating that the affect of summarisation may be increased for non-engaging full presentations.

6.3.4 Eye-Tracking Heat Maps

Heat maps are used to visualise the intensity of focus from eye-tracking participants for presentations. This is a graphical representation of the data where the intensity of focus on each pixel in the image is represented as a colour, from green to red. Pixels with no colouration indicate that there was no focus on this particular pixel by eye-tracking participants. The colours range from green to red, with green indicating little intensity of focus and red indicating high intensity of focus. For example, in Figures 6.2 and 6.3 we can see red blotches over the speaker, indicating that eye-tracking participants spent a large proportion of time focused on the presenter during this presentation.

Figures 6.2 and 6.3 show heat maps generated automatically from eye-tracking outputs. These show consistently more focus on speaker for full presentations and slightly more focus on slides for presentation summaries. While there are periods in which participants lose focus from the attention scene, this happens for too little time to show as significant in the heat maps. The higher focus on slides for presentation summaries demonstrates that participants spent a higher proportion of time reading through the slides for summaries than for full presentations, and also spend a higher proportion of time focusing on the speaker themselves during full presentations than for presentation summaries. This indicates that presentation summaries contain a higher concentration of new information than full presentations.

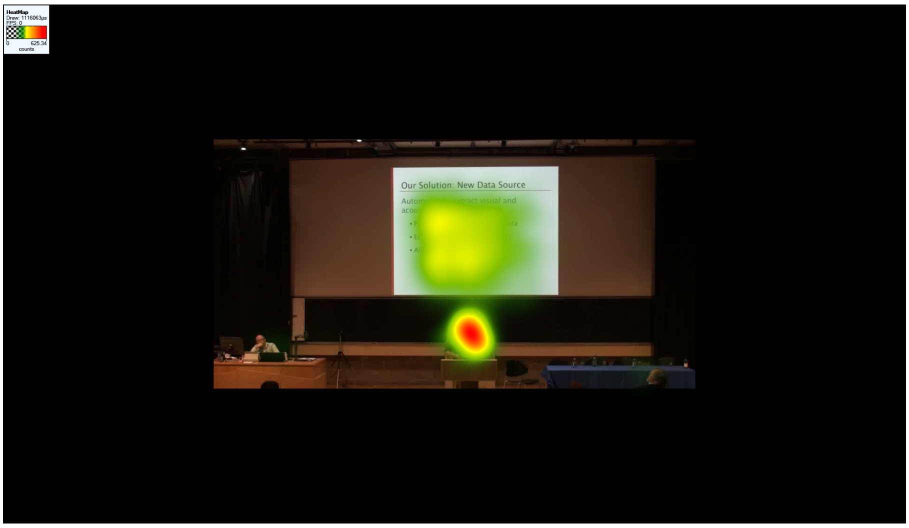


Figure 6.2: Pleanaryoral2 full - heat map 1

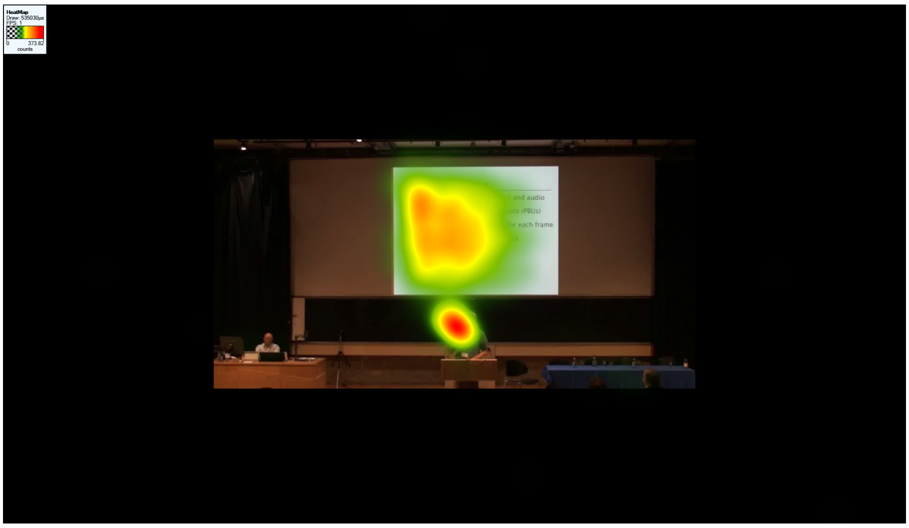


Figure 6.3: Pleanaryoral2 full - heat map 2

Figures 6.2 and 6.3 show heat maps for plenaryoral2 full version, with high intensity over the speaker and not too much intensity over the presentation slides. Figure 6.3 shows slightly more intensity over slides than Figure 6.2, both maps however show far more intensity over the speaker than over slides. Intensity on these maps indicates where eye-tracking participants have focused their gaze. This shows that participants held their gaze on the speaker for far longer periods than they spent reading the slides.

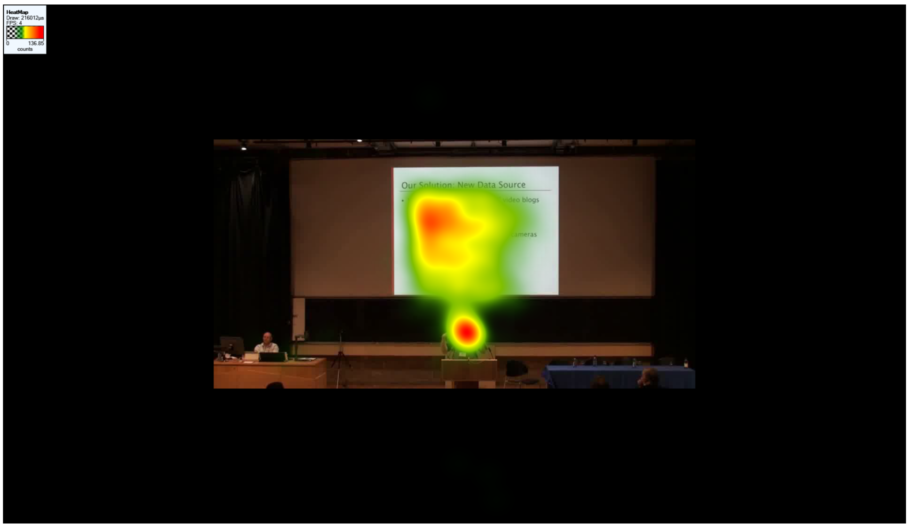


Figure 6.4: Plenaryoral2 summary - heat map 1

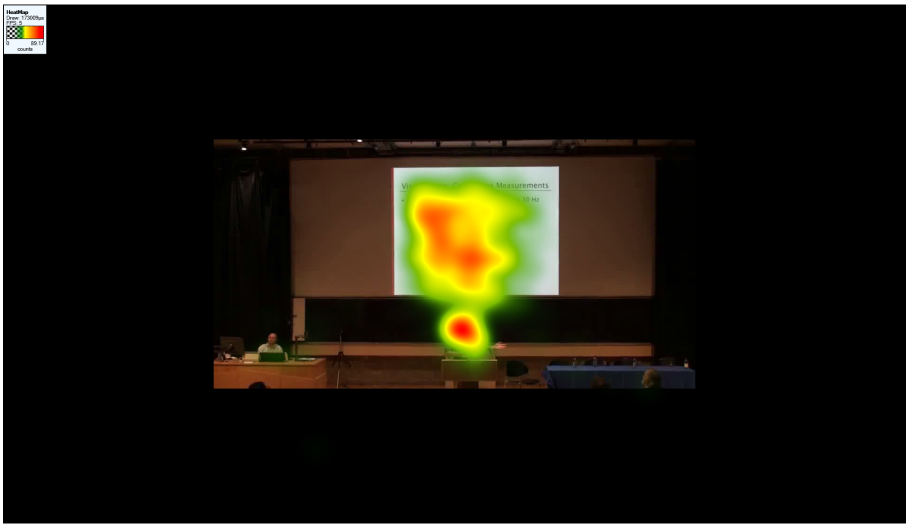


Figure 6.5: Plenaryoral2 summary - heat map 2

Figures 6.4 and 6.5 show heat maps from plenaryoral2 summary version, these show higher intensity over the presentation slides than heat maps for the full version. These maps do not show reduced intensity over the speaker, however with far more intensity over slides, almost equalling intensity over the speaker, this indicates that participants spend far more of their time reading the slides during the summaries than during full presentations. This indicates that participants were consuming new information for a higher proportion of the time when they were watching summaries than for full presentations.

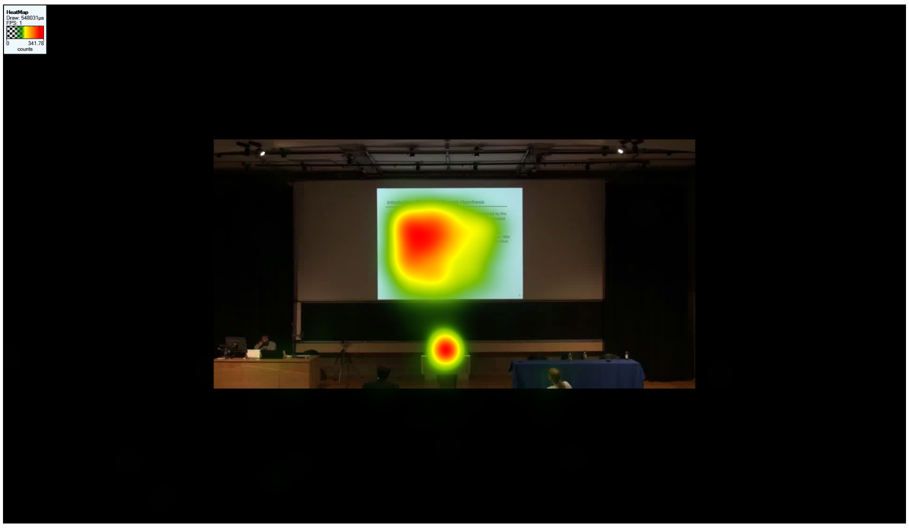


Figure 6.6: prp1 full - heat map 1

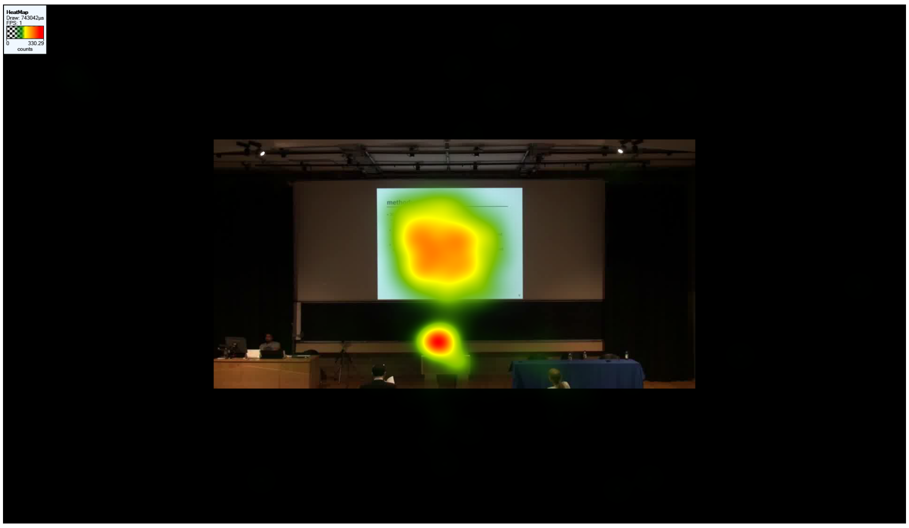


Figure 6.7: prp1 full - heat map 2

Figures 6.6 and 6.7 show heat maps from prp1 full version, and shows near equal intensity between speaker and presentation slides, indicating that participants spent near equal time focusing on slides and on the speaker. Figure 6.7 however shows slightly less intensity over slides than for speaker. This map also shows a slight tinge of intensity over the table to the speakers left, indicating that they have lost focus entirely over the attention area for small sections during these full presentations.

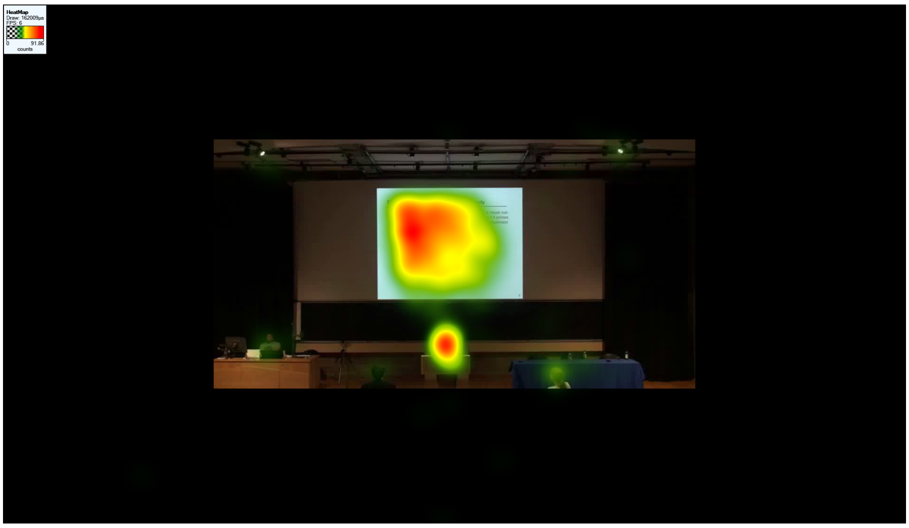


Figure 6.8: prp1 summary - heat map 1

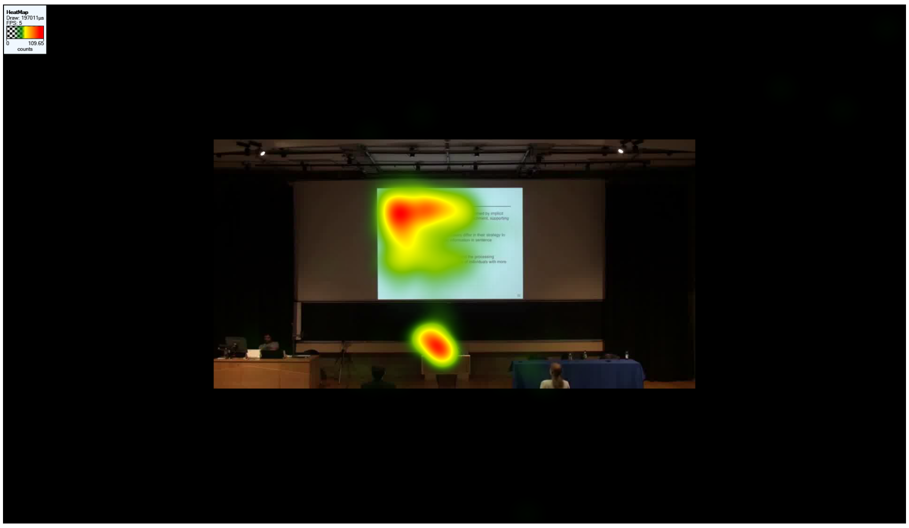


Figure 6.9: prp1 summary - heat map 2

Not too much difference is noticeable in the above heat maps of prp1, shown in Figures 6.8 and 6.9, compared to the previous heat maps of the full version of the same video, Figures 6.6 and 6.7. In Figure 6.8 however there are noticeable light blobs of heat around the edges of the stage, indicating some participants had a tendency to lose focus for short periods of time during these summaries, which suggests that these summaries were perhaps not as engaging as we would wish. In Figure 6.8 there is a noticeable increase of intensity over the presentation slides than for the corresponding image for the full version, indicating that participants spend more time reading the slides and taking in new information during the summary.

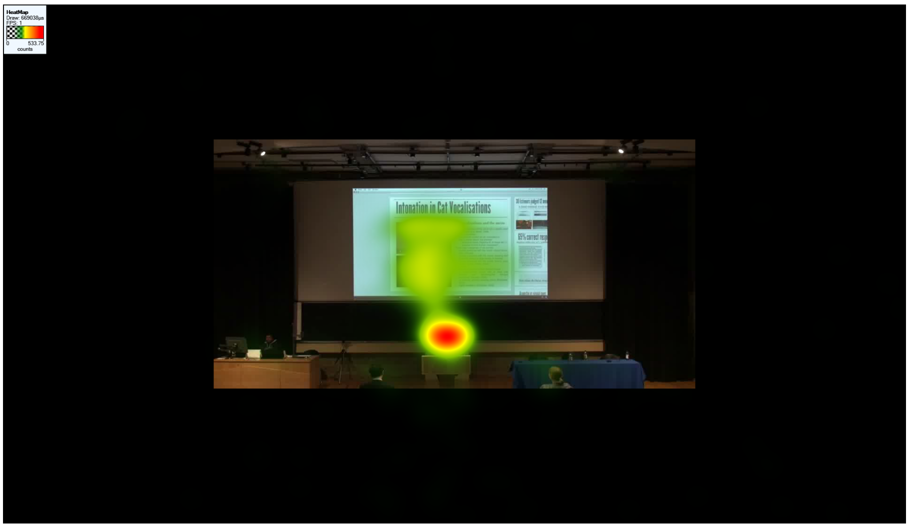


Figure 6.10: prp5 full - heat map 1

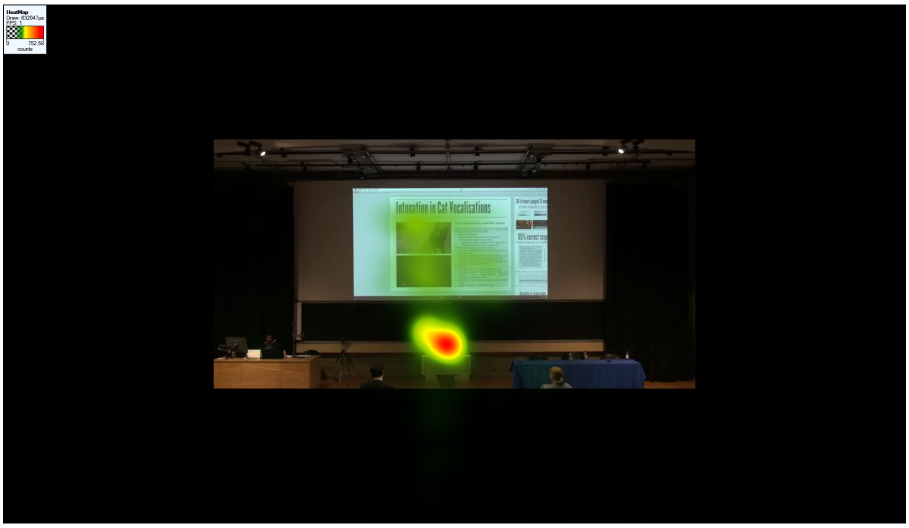


Figure 6.11: prp5 full - heat map 2

The heat maps from prp5 full version, shown in Figures 6.10 and 6.11, show that participants spent most of the time during this presentation focused on the speaker. Little intensity over the slides for these two maps indicate that little time was spent reading the slides for these presentations. It should be noted that this full presentation scored very highly for engagement.

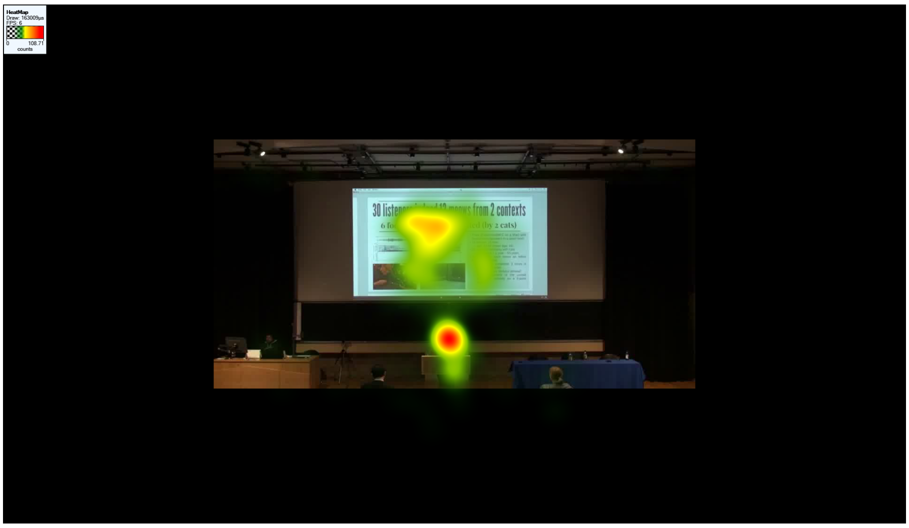


Figure 6.12: prp5 summary - heat map 1

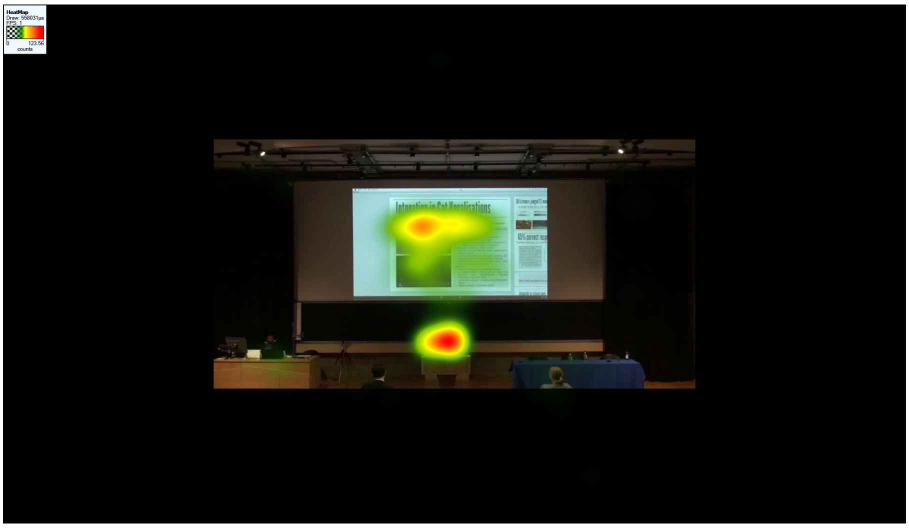


Figure 6.13: prp5 summary - heat map 2

Heat maps for prp5 summary version, Figures 6.12 and 6.13, show much higher intensity over the area of the presentation slides, indicating that participants spent a much higher proportion of time reading the slides during summaries than for the full presentation. Also noticeable is that this area of high intensity appears over the same part of the slides for both maps, Figures 6.12 and 6.13, with little covering the rest of the slides. It should be noted that slides for this presentation were designed more like newspaper clippings than a typical set of presentation slides. This presentation also scored very highly for audience engagement.

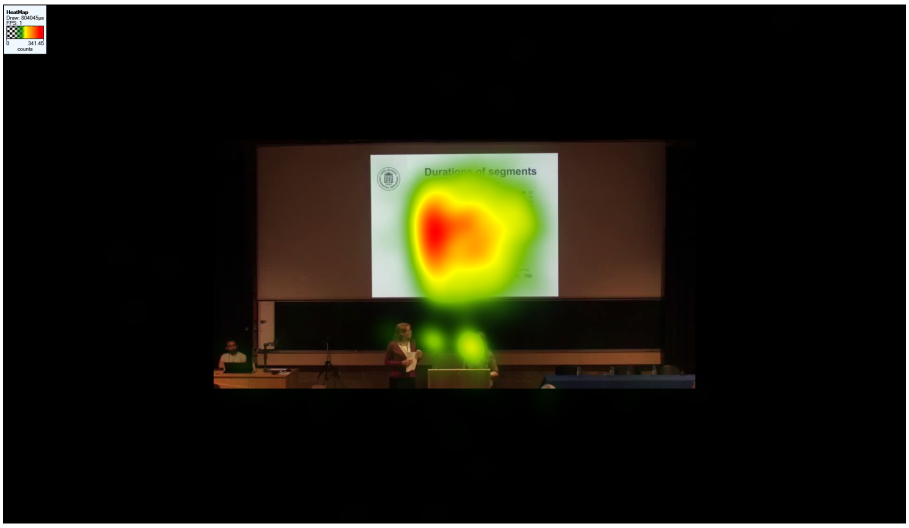


Figure 6.14: speechRT6 full - heat map 1

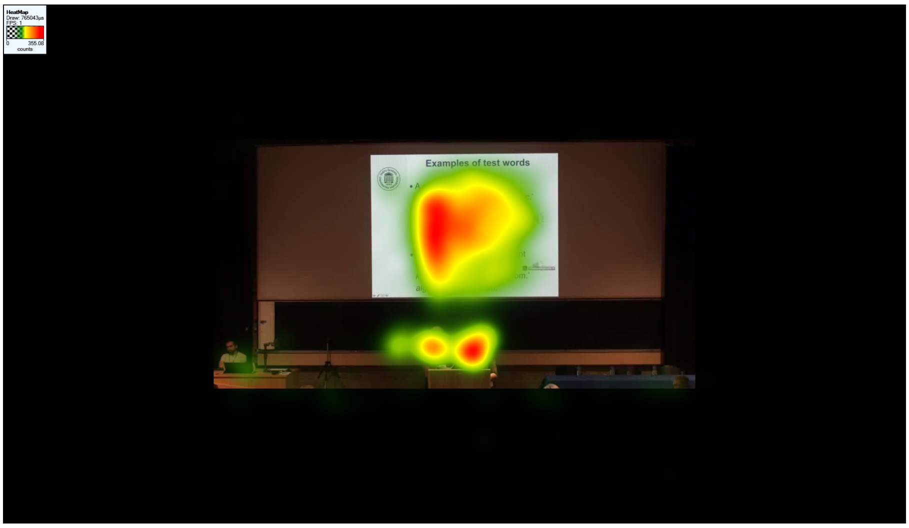


Figure 6.15: speechRT6 full - heat map 2

Heat maps above for speechRT6 full version, as shown in Figures 6.14 and 6.15, show that participants spent a very high proportion of the time looking at the presentation slides. In Figure 6.14, very little intensity appears over the speakers, indicating that participants spent the vast majority of their time looking at the slides. In Figure 6.15, light blobs are noticeable around the edges of the stage indicating that their were times during these full presentations in which participants lost focus. It should be noted that this presentation was found to be the least engaging presentation of the full dataset.

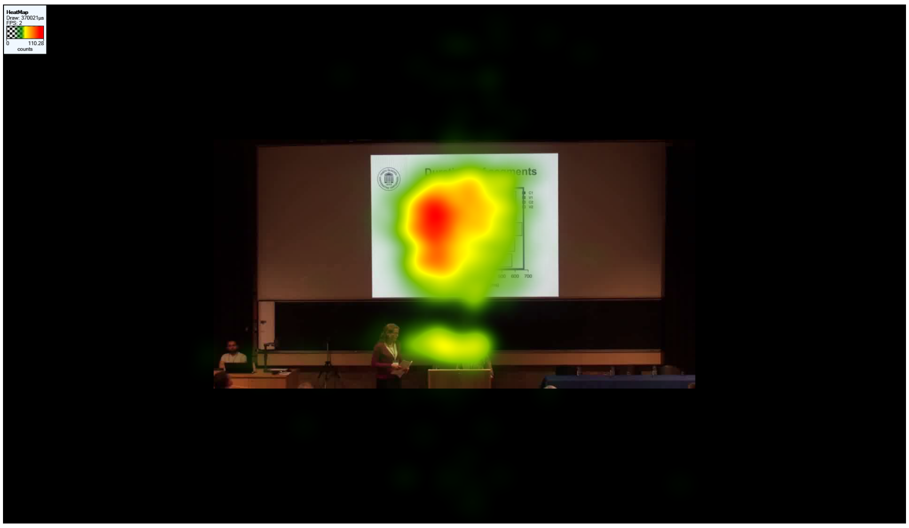


Figure 6.16: speechRT6 summary - heat map 1

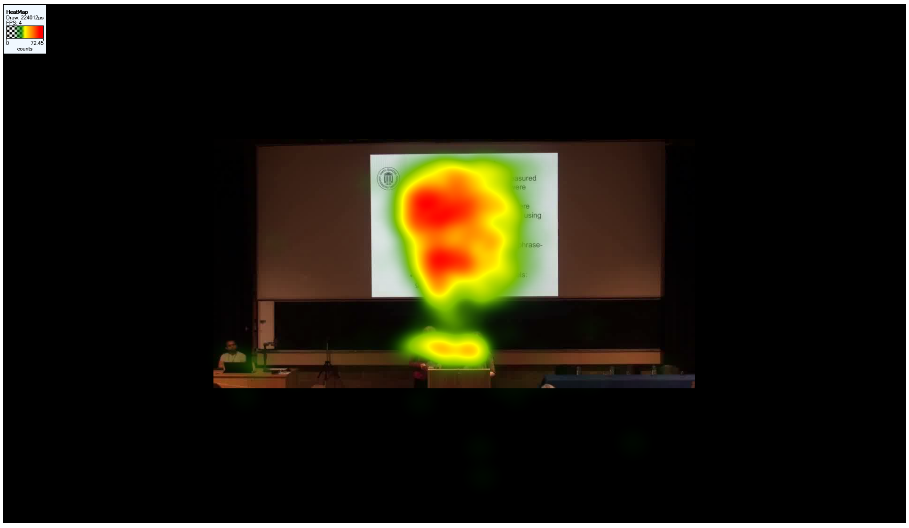


Figure 6.17: speechRT6 summary - heat map 2

Heat maps for speechRT6 summary version are shown in Figures 6.16 and 6.17, and show an even higher intensity over presentation slides, and even less intensity over the speakers. Intensity in Figure 6.17 is also much more spread around the whole area of the slides, indicating that participants spend almost the entire time during these summaries reading the slides to consume the new information. There are also less blobs around the edges indicating that participants lost focus less frequently during summaries than for full presentations.

6.3.5 Questionnaire for Eye-Tracking Participants

Following participants completion of the session of video viewing with eye-tracking, they completed a questionnaire. This questionnaire asked them how much they agreed with each of 5 statements, depending on whether they had watched a summary or the full video, on a 5-point Likert scale.

The five statements on each video are shown in Table 6.14, with the available levels of agreement shown in Table 6.15.

Table 6.14: Statements ranked by participants

#	Statement
1	This summary is easy to understand.
-	This video is easy to understand.
2	This summary is enjoyable.
-	This video is enjoyable.
3	This summary is informative.
-	This video is informative.
4	This summary is coherent.
-	This video is coherent.
5	This summary would aid me in deciding whether to watch the full video.
-	I would have preferred to see a summary of this video.

Table 6.15: Levels of Agreement

#	Level of Agreement
1	Strongly Disagree.
2	Disagree.
3	Neutral.
4	Agree.
5	Strongly Agree.

Tables 6.16 and 6.17 show the average rankings for each video based on answers given to the questionnaire.

1. Video summaries were rated as being slightly less easy to understand than full presentation videos, with the largest different being for video 2, prp1.

2. All video summaries have also received high ratings for their ability to help users to decide whether they wished to watch the full presentation video.

3. Summarisation videos all received good scores for informativeness, with all videos receiving a score of over 2.5. The best scoring was again video 2, prp 1, which received an overall informative ranking of 3.22. The video found to be the least informative was video 4, speechRT6. This video was chosen for summarisation evaluation as it had been found by our classifiers to be the least engaging and least comprehensible presentation from the dataset. This video summary still received an overall informativeness ranking of 2.5 out of 5, indicating that the summarisation strategy maintains informativeness of presentation videos when generating summaries.

4. Three of the four summaries evaluated were found to be slightly less enjoyable than the original video, indicating that presentations do tend to lose some enjoyableness after summarisation. The exception to this is video 4, speechRT6. As explained previously this video was found by our classifiers to be the least engaging and least comprehensible video from the collection. That this video summary is found to be more enjoyable than the corresponding original video indicates that watching summaries can be a good alternative to watching full presentations in which the user may not be very interested, or in cases where the presentation in question is hard to follow or simply not very engaging.

It should be noted that none of the results shown in Tables 6.16 and 6.17 below are statistically significant, limiting the conclusions which can be drawn from such results.

Table 6.16: Averaged rankings per video

Video	Q1-avg	Q2-avg	Q3-avg	Q4-avg	Q5-avg
video-1 summ	1.750	1.750	2.625	2.625	3.750
video-1 full	2.500	2.750	4.000	3.500	3.750
video-2 summ	2.889	2.889	3.222	2.889	4.333
video-2 full	4.500	4.000	3.700	4.100	4.100
video-3 summ	3.000	3.250	3.125	3.500	4.125
video-3 full	4.125	3.625	4.500	4.500	3.125
video-4 summ	1.700	1.800	2.500	2.400	3.800
video-4 full	2.333	1.667	3.333	3.333	4.222

Table 6.17: Averaged rankings per summary or full video

Video Type	Q1-avg	Q2-avg	Q3-avg	Q4-avg	Q5-avg
Generated Summary	2.314	2.400	2.857	2.829	4.000
Full Presentation	3.400	3.029	3.857	3.857	3.829

As can be observed in Tables 6.16 and 6.17, average coherence rankings are fund to be 2.8, indicating that the summarised videos maintain their coherence. The video summaries also receive better scores than the full presentation videos for Question 5, at 4 out of 5, indicating that summarised videos are very helpful for aiding users to decide if they wished to watch the full presentation video.

6.3.6 Evaluation by Comparison with Enhanced Digital Video Browser

For the final evaluation of presentation summaries, we perform a comparison against the properties of an enhanced digital video browser (Li et al., 2000). In this work, the authors evaluated the effectiveness and most useful features of an enhanced digital video browser. Within the domain of conference presentations they found that Time Compression and Pause Removal were rated by participants as the most useful features, and the enhanced digital video browser was found to allow users to watch substantially more of a presentation than time would typically allow for. For further evaluation of our automatically generated presentation summaries, we compare against the use of an enhanced digital video browser providing the features of Pause Removal and Time Compression.

For these evaluations, participants were given 5 minutes to gain a clear and concise overview of a presentation which they had not seen before, in order to be able to take part in a meeting discussing what was presented. The enhanced digital video browser removed pauses and allowed participants to increase or decrease playback rate as desired, up to a maximum of 2.5 times normal speed. Using this, the participants had 5 minutes to gain a clear and concise overview of one of the presentations, and following this, answered 3 questions on the use of the browser.

Participants also were given an automatically generated summary of the presentation for the same purpose. This summary was under 5 minutes in length and participants were again given 3 questions to answer on the use of the browser. The purpose of this investigation was to examine whether summaries or the enhanced video browser is better for gaining a rapid concise overview of presentations where users do not have the time to watch the full video.

Table 6.18 shows the three statements relating to each tool for which participants indicated their level of agreement on the 7-point Likert scale shown in Table 6.19:

Table 6.18: Statements ranked by participants

#	Statement
1	This tool is easy to use.
2	This tool allowed me to gain a clear and concise overview of the presentation.
3	This would be my choice for tasks of this nature.

Table 6.19: Levels of Agreement

#	Level of Agreement
1	Very Much Disagree.
2	Disagree.
3	Disagree Somewhat.
4	Neutral.
5	Agree Somewhat.
6	Agree.
7	Very Much Agree.

A total of 4 videos from the dataset of 31 presentations were selected for evaluation. These include video's which were previously rated by our human annotators as most engaging, most comprehensible, best speaker ratings and the least engaging video.

This evaluation was separated into 8 different tasks - Each of the four videos chosen for evaluation was associated with 2 tasks, one beginning with the enhanced digital video browser and the other task beginning with the presentation summary. 8 participants were recruited for each task, totalling 64 participants in total.

Table 6.20: Averaged ratings per video

Video	Q1-avg	Q2-avg	Q3-avg
video-1 summ	4.563	4.375	3.625
video-1 enh	4.500	3.938	3.938
video-2 summ	4.750	4.688	4.188
video-2 enh	5.125	5.188	4.750
video-3 summ	4.875	4.688	4.750
video-3 enh	4.125	4.313	4.375
video-4 summ	4.813	4.000	4.000
video-4 enh	4.438	4.063	3.938

Table 6.21: Total ratings per type

Q	Summary	Enhanced
1	304	291
	4.750	4.547
2	284	280
	4.438	4.375
3	265	272
	4.141	4.250

Tables 6.20 and 6.21 show ratings for participant comparison of the summaries and the video browser. The results show that participants are 4.5% more likely to find that summaries are easier to use than the enhanced digital video browser for gaining a quick, clear and concise overview of missed presentations. Participants are also 1.5% more likely to agree that the automatically generated video summaries are better for gaining a clear and concise overview of missed presentations than the use of the enhanced digital video browser. However, they are still 2.6% more likely to choose the enhanced digital video browser as their tool of choice.

It must be pointed out that these results are again not statistically significant.

6.3.7 Evaluation Between Summary Types

Table 6.24 shows a comparison between summaries built using all available features as described in this thesis, and summaries built using only the important keywords extracted from the ASR transcripts. Summaries built using only keyword information do not include any classified paralinguistic features, and simply include sentences containing the most keywords.

Table 6.25 shows further comparisons between summaries built using all available features, and summaries built using only a subset of features. For audio-only summaries, classification of the paralinguistic features of Speaker Ratings, Audience Engagement, Emphasis, and Comprehension was performed as described in the earlier chapters, but only the audio features were used to determine which information to include in summaries. Similarly, for visual only summaries, classification to select content for inclusion in the summaries was performed using only visual features. For no keyword summaries, classification of content for inclusion in the summaries was performed based on all audio and visual features excluding only the keywords.

Table 6.24 shows results of the comparison between summaries built using keyword information only and summaries built using all available information. A total of 48 participants watched the summaries and answered the questionnaire on each summary. The questions addressed in this comparison are shown in Table 6.22, and answers were ranked on a Likert scale from 1 to 7, as shown in Table 6.23.

Table 6.22: Statements rated a Likert scale by participants

#	Statement
1	This summary is easy to understand.
2	This summary is informative.
3	This summary is enjoyable.
4	This summary is coherent.
5	This summary would aid me in deciding whether to watch the full video.

Table 6.23: Likert scale Levels of Agreement to statements in Table 6.22

#	Level of Agreement
1	Very Much Disagree.
2	Disagree.
3	Disagree Somewhat.
4	Neutral.
5	Agree Somewhat.
6	Agree.
7	Very Much Agree.

The results in Table 6.24 indicate that participants perceive little difference

Table 6.24: Likert scale level of agreement per summary type for each video

Video	Q1	Q2	Q3	Q4	Q5
plen2_Key	4	5	4	5	4
plen2_All	4	5	4	4	4
prp1_Key	3	5	3	4	3
prp1_All	3	5	3	4	3
prp5_Key	5	6	5	5	5
prp5_All	4	6	5	5	5
spRT6_Key	4	5	3	3	3
spRT6_All	3	5	3	4	3

overall between summaries generated using keyword information only, and those generated using all available information. The most notable results indicate that users perceive the keyword summary of Video 3, prp_5, to be easier to understand than the summary using all available information for this video. Interestingly, they also perceive the summary generated using all available information from Video 4, speechRT6, video to be much more coherent than summaries generated using only keyword information from the corresponding video. This may be explained by the fact that speechRT6 was labelled by our classifiers as less engaging and comprehensible than other videos in the collection, while prp_5 was found to be the most engaging video. This would indicate that the effectiveness of using engagement information in the generation of summaries depends on just how engaging the video was in the first place.

Table 6.25 shows results of summaries generated using audio-only classifications in combination with keywords, visual-only classifications in combination with keywords, audio and visual classifications in combination with keywords, and audio and visual classifications with no keywords. Audio-only classification summaries and visual-only summaries are both rated slightly less easy to understand and informative than summaries built using all features without keywords. Summaries built using no keywords also lack coherence, while nearly all summaries score highly on whether they would help users decide if they wanted to see the full presentation.

Table 6.25: Likert scale level of agreement per summary type for each video

Video	Q1	Q2	Q3	Q4	Q5
plen2_Classify	2.625	3.750	3.125	3.438	4.625
plen2_aud_only	2.313	3.500	2.438	3.813	4.813
plen2_vid_only	3.000	4.625	2.438	4.063	4.750
plen2_no_key	3.563	4.813	3.750	4.250	5.063
prp1_Classify	2.875	4.313	2.313	3.813	4.500
prp1_aud_only	2.250	3.875	2.438	2.938	4.938
prp1_vid_only	2.375	3.250	2.000	2.875	4.063
prp1_no_key	2.563	3.813	2.813	3.875	5.063
prp5_Classify	4.250	4.875	4.500	4.438	4.813
prp5_aud_only	3.625	4.250	3.375	3.625	5.125
prp5_vid_only	4.250	4.500	4.438	4.125	5.313
prp5_no_key	5.063	5.063	3.813	4.563	4.875
spRT6_Classify	2.875	4.125	2.625	3.875	5.438
spRT6_aud_only	2.813	4.563	2.500	4.000	5.063
spRT6_vid_only	2.000	3.500	2.500	3.063	5.125
spRT6_no_key	2.688	3.938	2.438	3.313	4.500

6.3.8 Gaze Plots

In this section, we examine gaze plots from eye-tracking experiments. By looking carefully at plots for full and summary videos, the difference in attention and focus for different video types becomes more clearly defined. Gaze plots are data visualisations which can communicate important aspects of visual behaviour clearly. The following sub-section explains how to interpret gaze plots.

6.3.8.1 Interpreting Gaze Plots

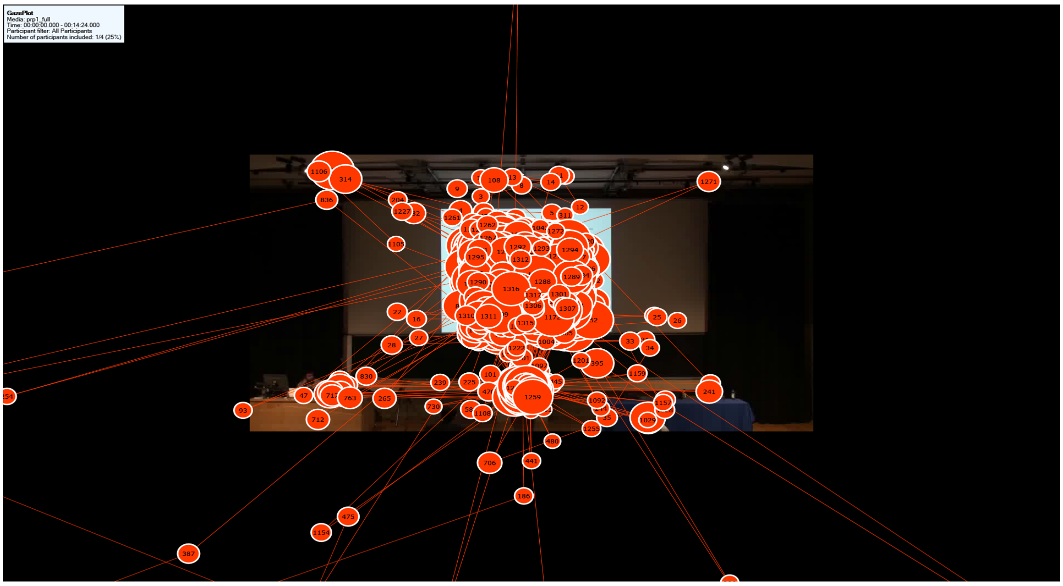


Figure 6.18: Example Gaze Plot for Interpretation

Gaze plots show the order, location and time spent looking at different areas of the screen. The number at the centre of the circle is the order, for example a circle with 1 at its centre indicates that that was the first part of the screen the participant gazed at, followed by 2, 3, etc. The location of each circle is a point of the participants gaze. The diameter of the circle indicates the length of time spend fixating at that point. For example, circle *a*, having twice the diameter of circle *b*, means that the participant spent twice as long fixating at point *a* than at point *b*.

From the example plot in Figure 6.18, we can see some gaze locations outside of the main attention area around the frame which are quite small. This indicates that the participant may have lost focus during these times but just for short periods. We can see quite a lot of circles, indicating a lot of fixations, over the speaker and over the presentation slides. The circles over the speaker are quite large meaning that these fixations were quite long. While are also large fixations over the slides, many of these fixations are quite short, which may be because the participant was reading over the slides. Further to this, previous work by (Rayner and Sereno, 1994) has shown that a large number of shorter fixations is consistent with higher cognitive activity.

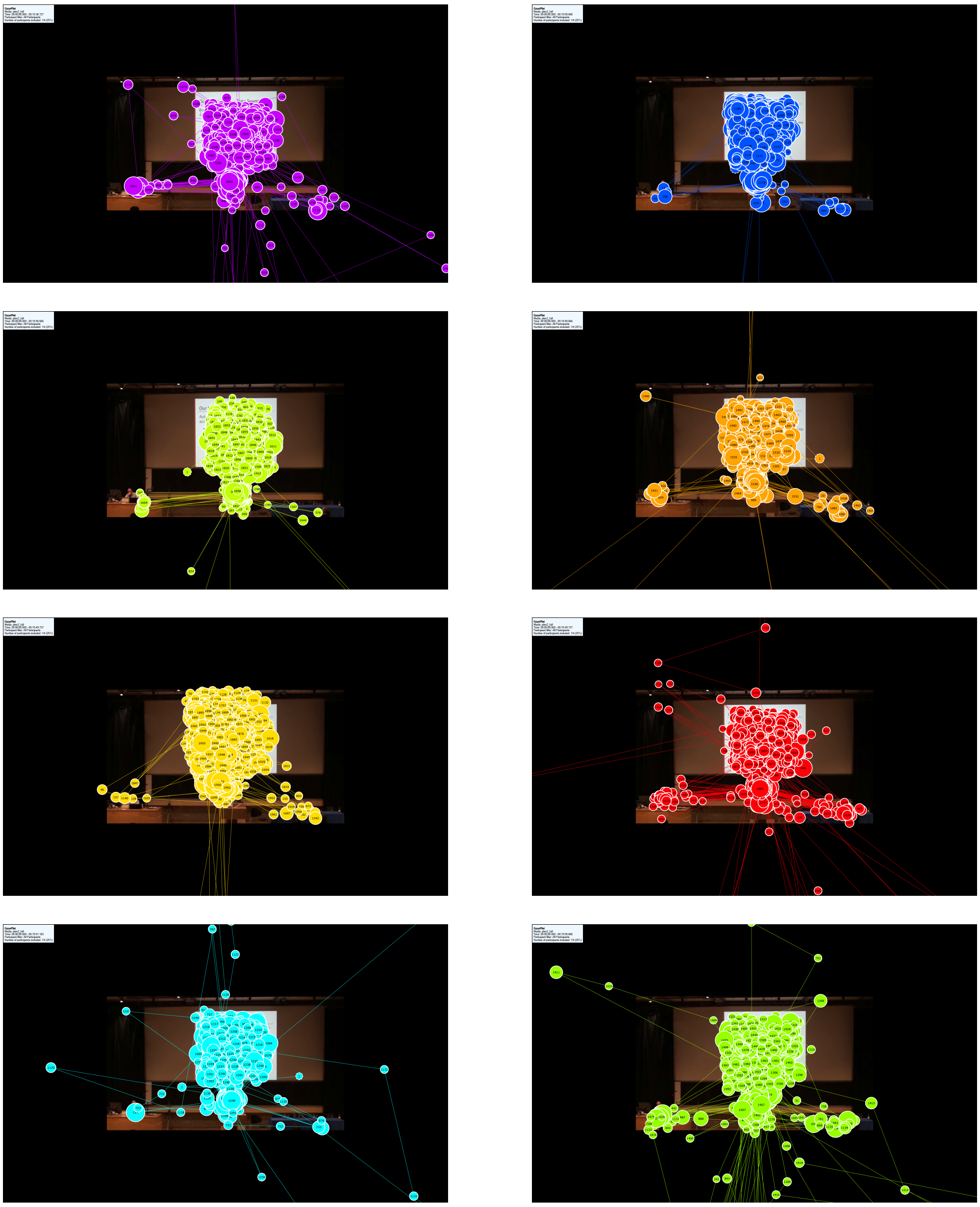


Figure 6.19: Plen 2 Full - Gaze Plots

The gaze plots shown in Figure 6.19, for plen2 (Video 1) full version, show that participants hold a quite a high level of attention with the full version of this presentation. Despite participants showing a tendency to switch attention momentarily to either bottom corner of the stage, the vast majority of the time participants are fully focussed on either the presentation slides or on the speaker themselves.

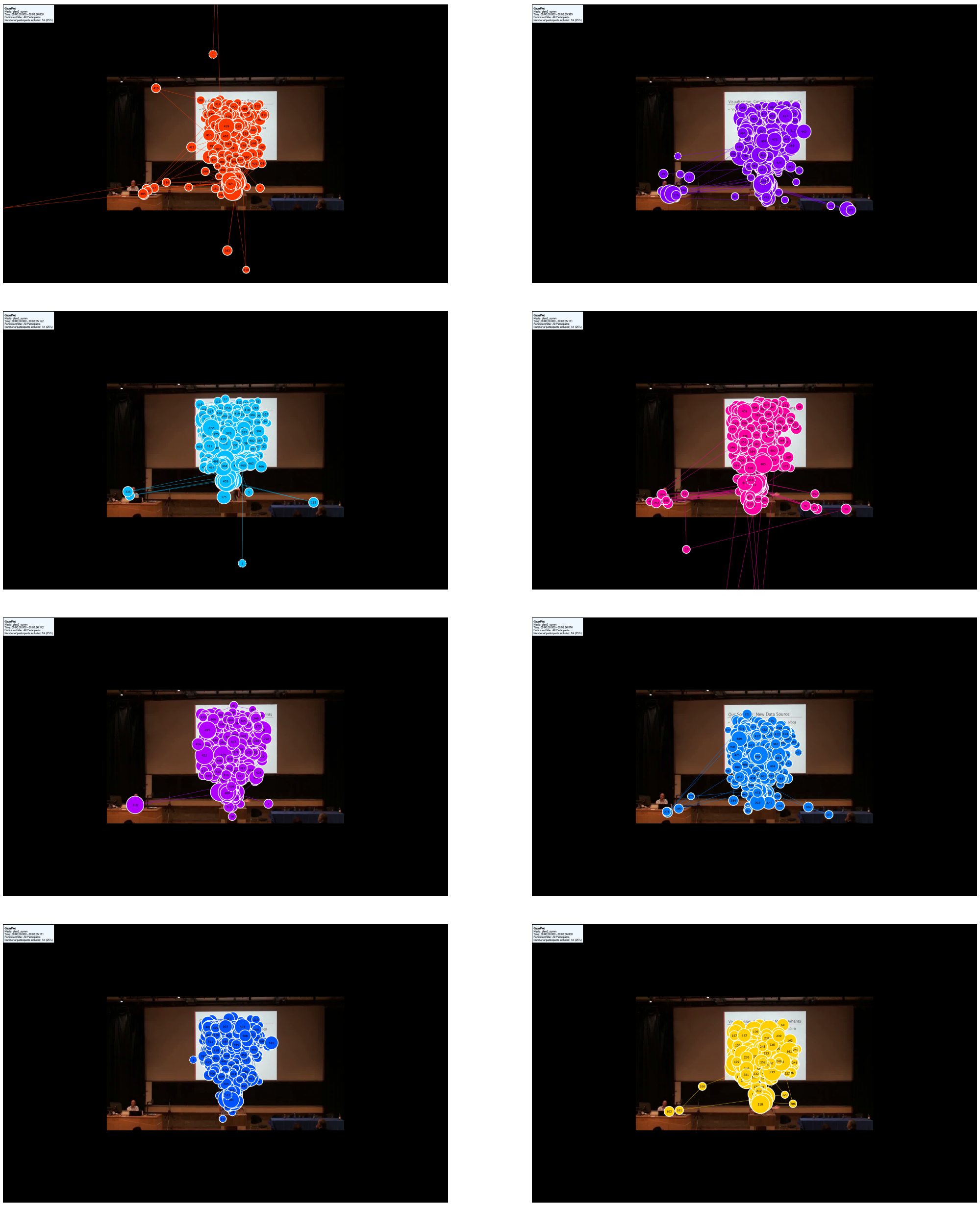


Figure 6.20: Plen 2 Summary - Gaze Plots

From the gaze plots in Figure 6.20, plen2 summary version, we can see that even though participants already held a high level of attention during the full presentation for this video, this is improved further for the summaries, with a smaller proportion of participants spending time switching attention to either corner of the stage, and high levels of attention on the slides and on the speaker.

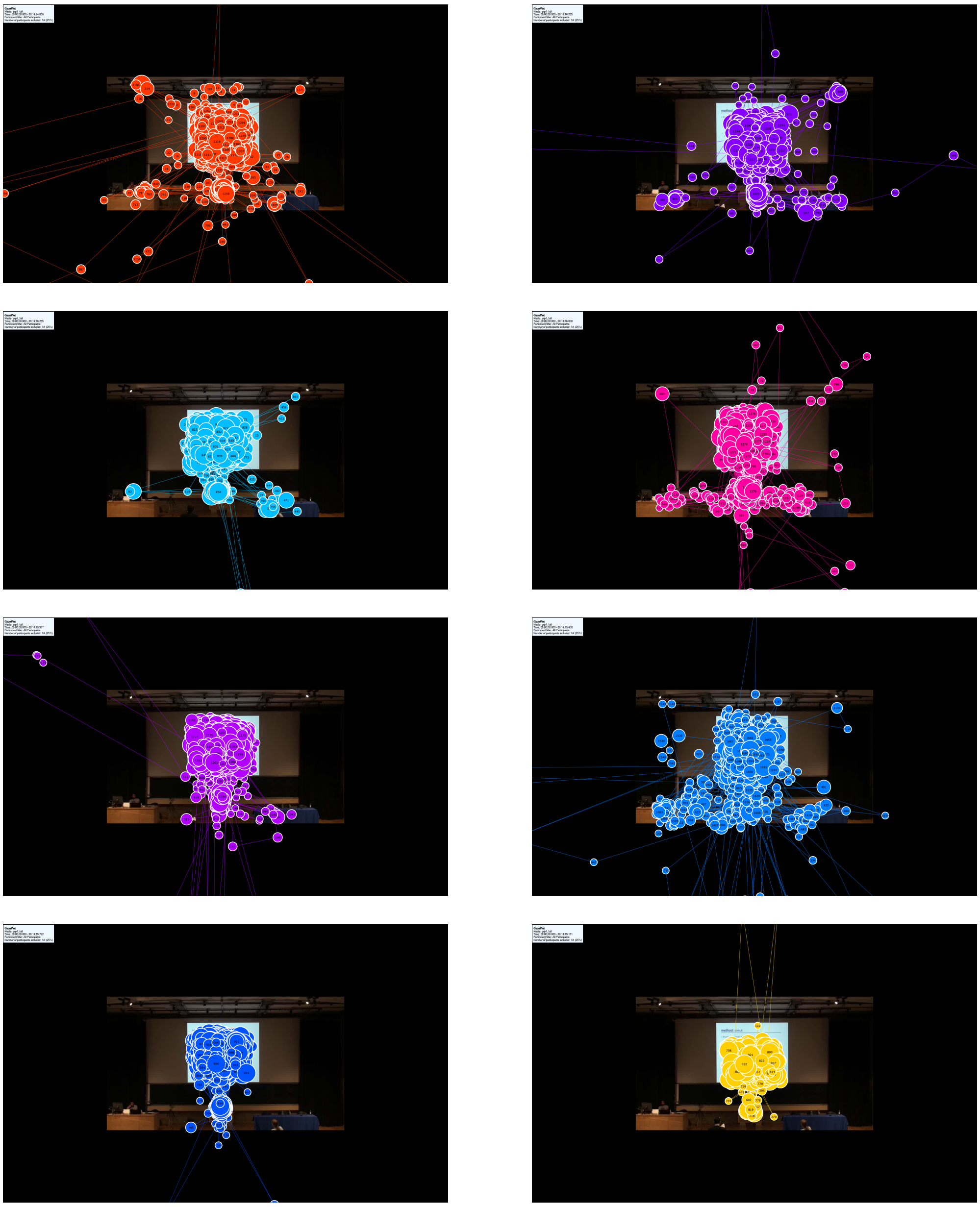


Figure 6.21: Prp 1 Full - Gaze Plots

The gaze plots in Figure 6.21, prp1 (Video 2) full version, show that around 50% of participants maintain a very high level of attention during this presentation, with nearly all of the focus being on the presentation slides and on the speaker, with very small proportions of time spend losing attention. Other participants maintain a fairly high level of attention throughout, but with more instances of them losing attention from the speaker and slides.

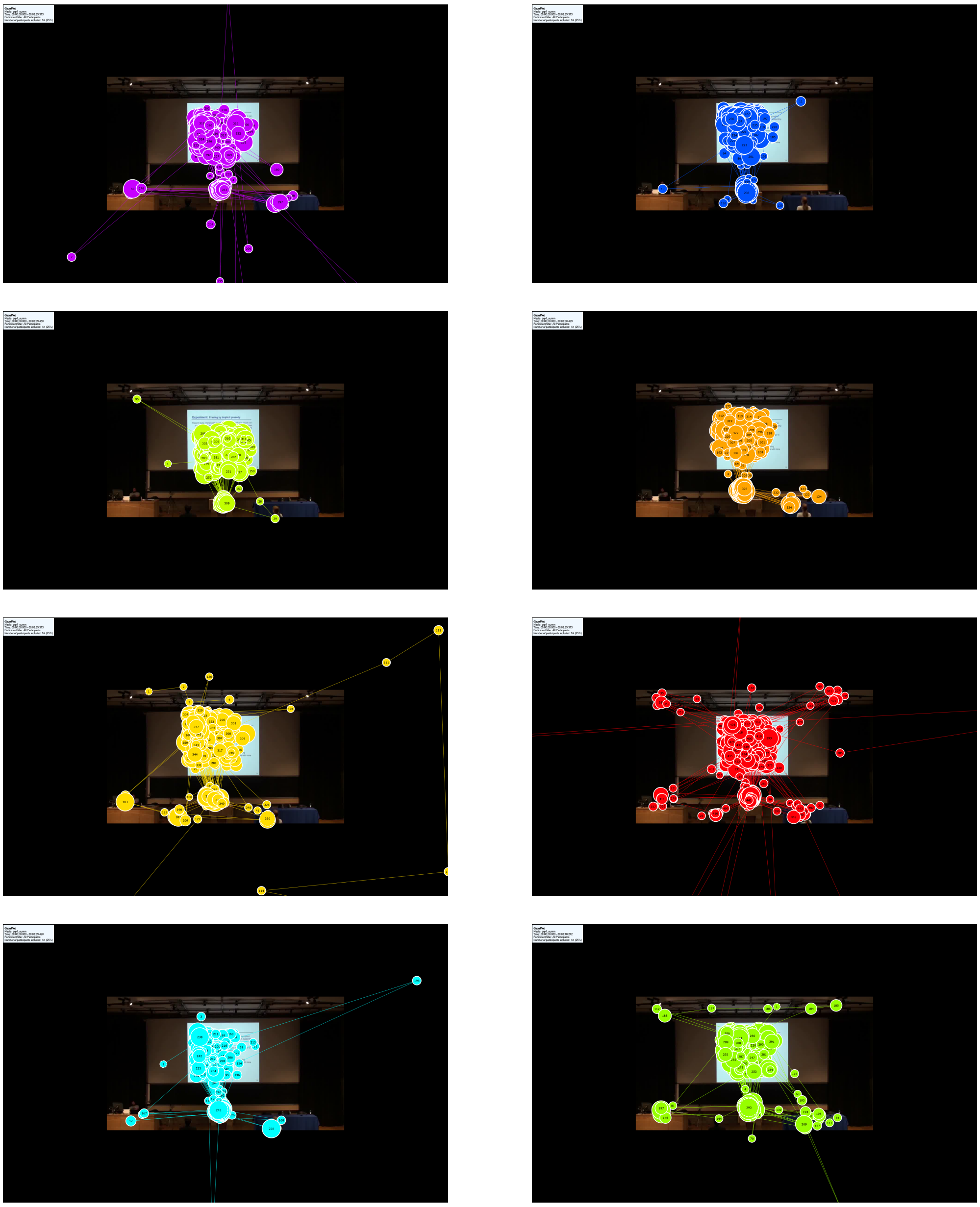


Figure 6.22: Prp 1 Summary - Gaze Plots

In Figure 6.22, prp1 summary version, we once again observe very high levels of attention for about 50% of participants, though these do not show improvement from the full version plots. Around another 50% maintain high levels of engagement and these show a clear improvements from the low-attention participants plots observed in Figure 6.21.

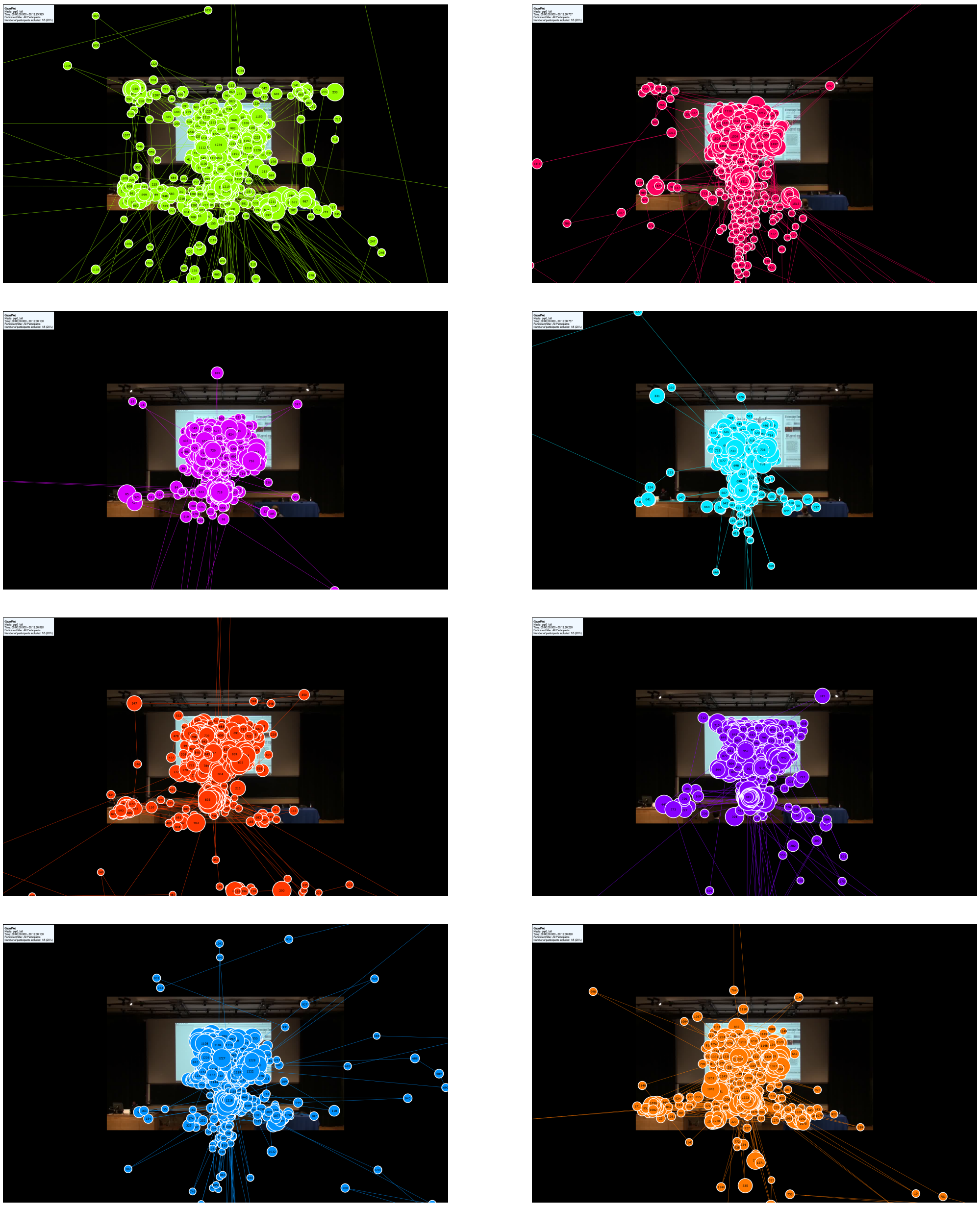


Figure 6.23: Prp 5 Full - Gaze Plots

Despite having being observed by human annotators as being the single most engaging presentation, The gaze plots in Figure 6.23, shows that prp5 (Video 3) full version maintains quite a low level of attention on the slides and speaker, with one participant in particular frequently switching their attention all around the scene. Other participants also show a tendency to frequently switch their attention all around the scene, to one spot at the lower left corner of the stage in particular, though this is slightly less refined than for the first participant.

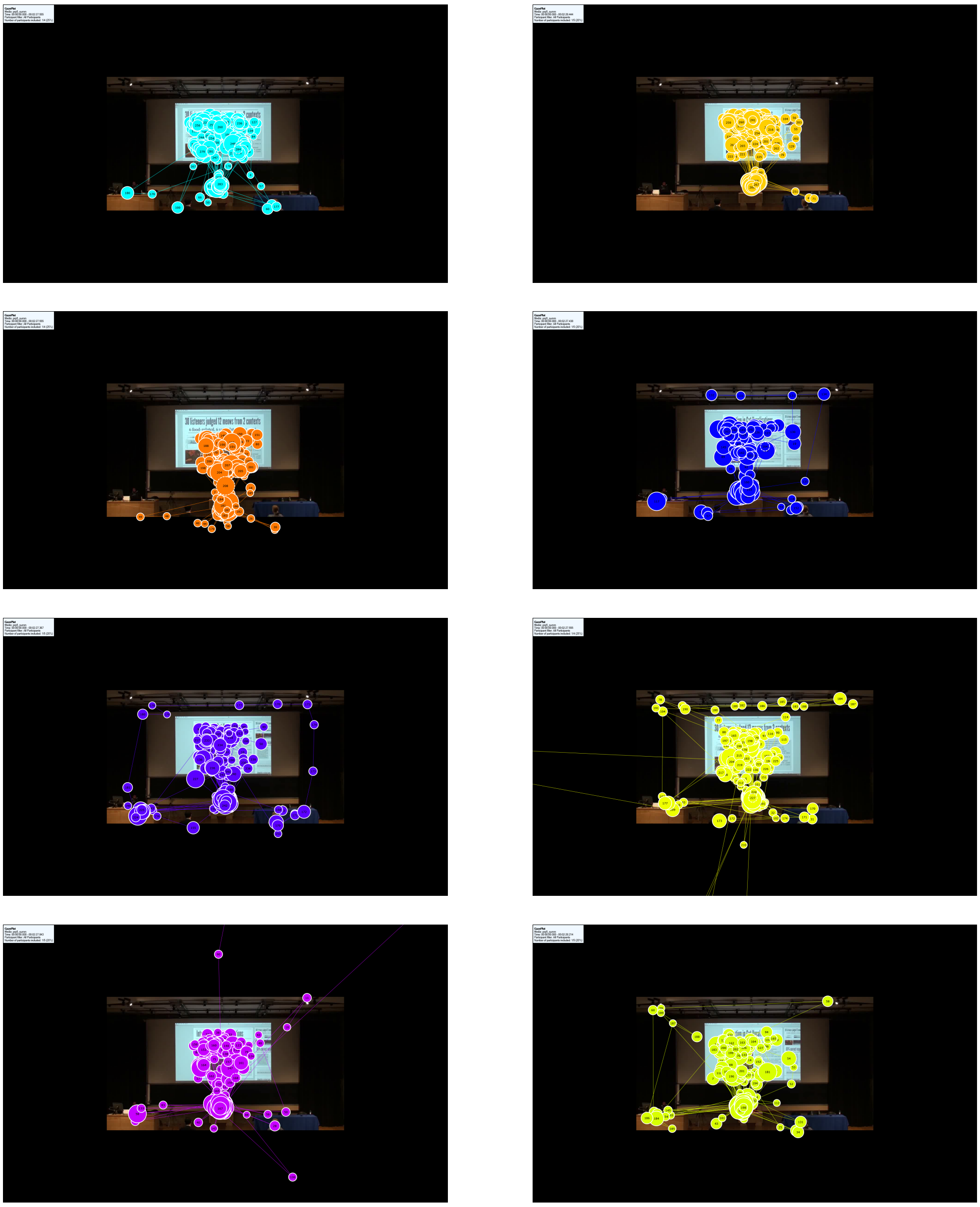


Figure 6.24: Prp 5 Summary - Gaze Plots

A big improvement on attention levels of participants for prp5 summaries can be seen in Figure 6.24. Here, participants are far more focussed on the slides and on the speaker during summaries. These plots also indicate that when participants did tend to switch their attention momentarily during summaries, they tended to switch their focus to defined areas outside of the attention scene of the slides and the presenter. Given that this presentation had a smaller amount of readable information on slides than others, with the majority of information existing in the audio stream, this could potentially be a case of participants listening carefully to the audio stream.

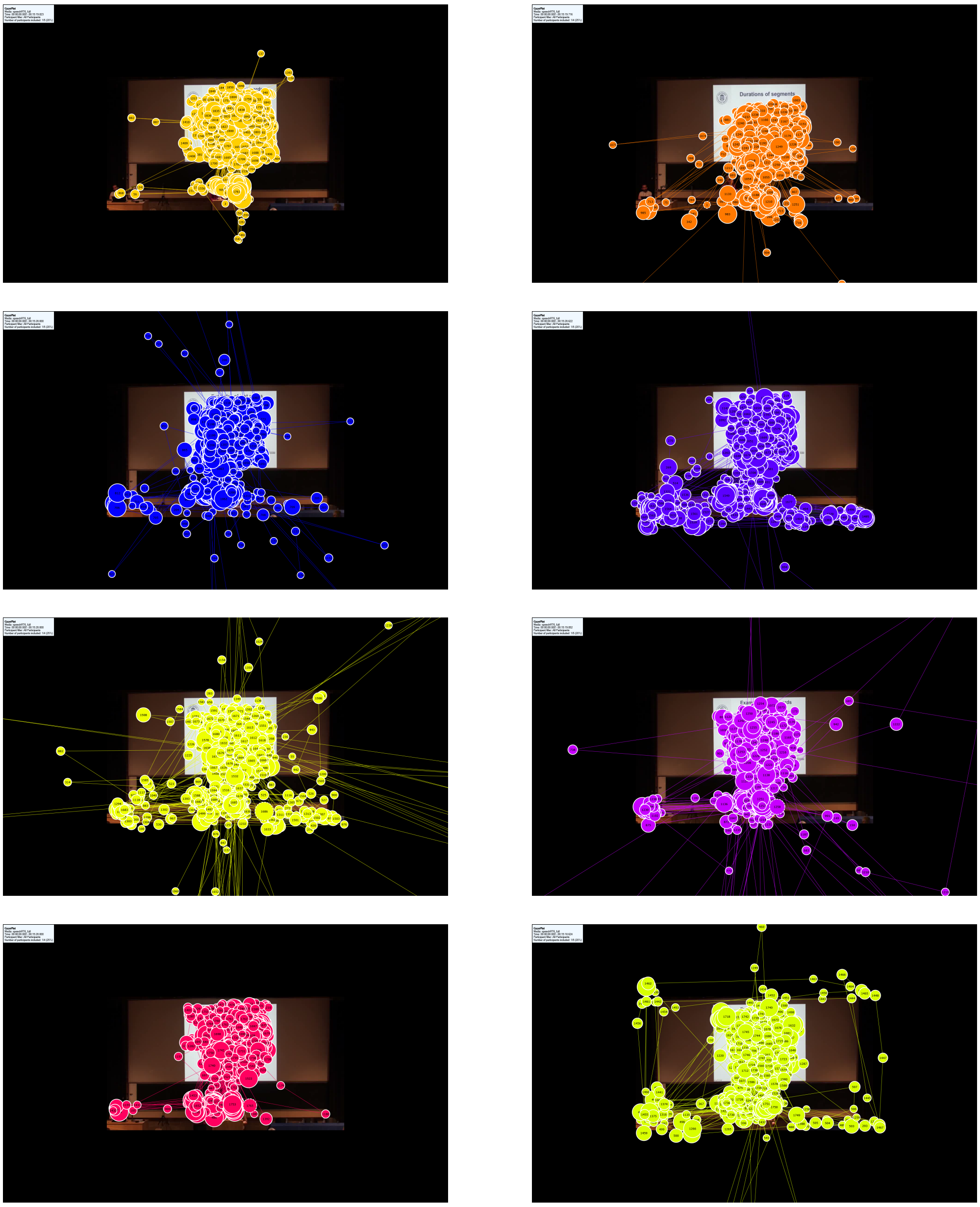


Figure 6.25: SpeechRT 6 Full - Gaze Plots

The gaze plots in Figure 6.25, speechRT6 (Video 4) full version, show that while two participants maintain quite high levels of attention during the full presentation (top left and bottom left plots), all others have trouble maintaining attention levels. This can be explained by the fact that this presentation was found by annotators to be the least engaging presentation, given by quite poor presenters.

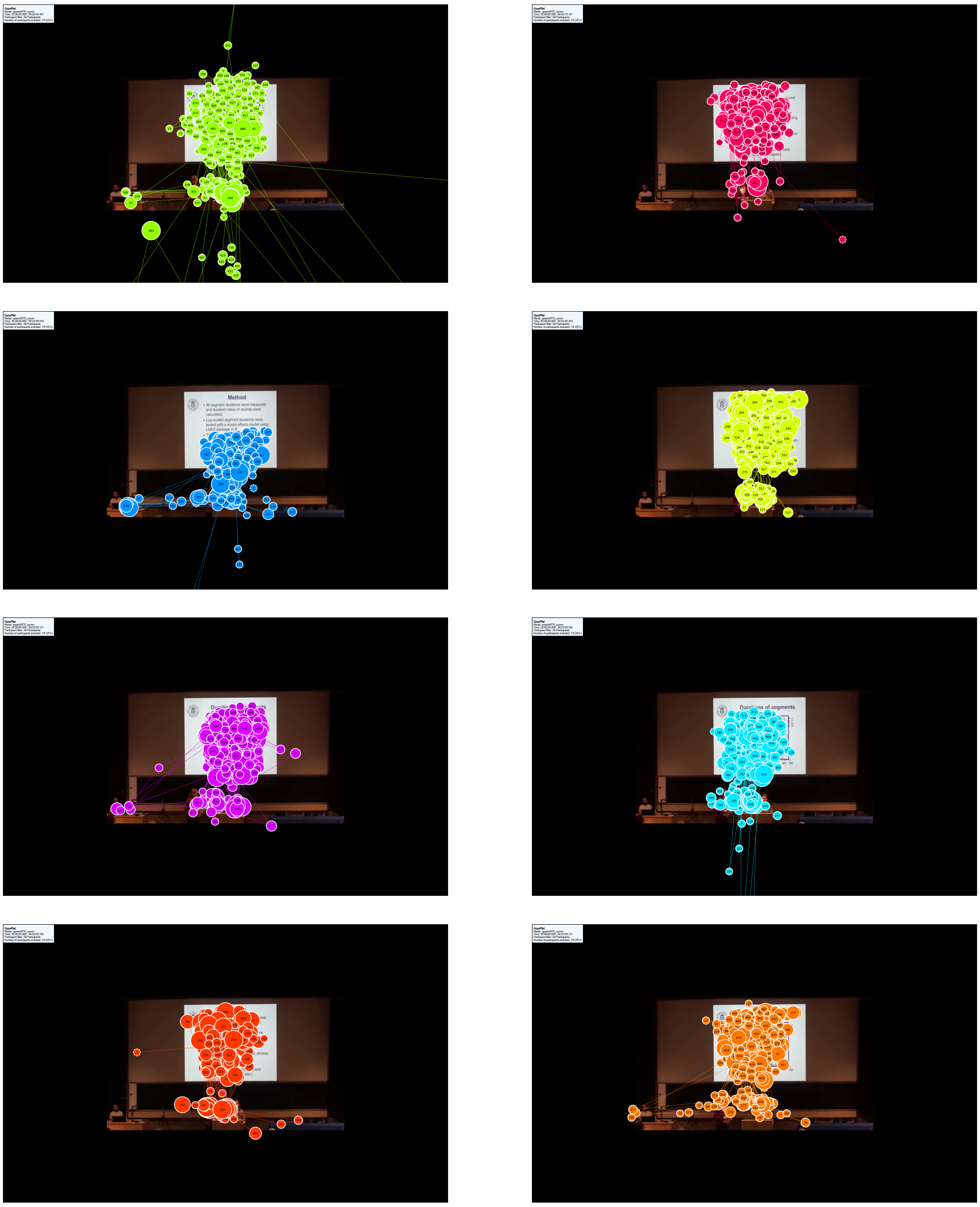


Figure 6.26: SpeechRT 6 Summary - Gaze Plots

As can be observed in Figure 6.26, speechRT6 summaries, participants maintain far higher levels of attention during summaries than during full presentations, with participants rarely switching their attention from the slides and the speaker. The full affect of this summarisation strategy can best be observed here by comparing full presentation plots in Figure 6.25 with summary plots in Figure 6.26, with full presentations having previously being found to be very disengaging, summaries show big improvements in participants attention levels over full presentations.

6.4 Summary

In this chapter we presented the summarisation algorithm used to create summaries of academic presentations using the extracted paralinguistic features from the previous chapters of this thesis. A number of evaluation strategies were employed, including eye-tracking, crowdsourced questionnaires and a crowdsourced comparison against the use of an enhanced digital video browser.

Summaries were built using all available features, and further summaries built using a subsection of all available features were compared. Eye-tracking compared the slides area of the scene and the speaker area of the scene. Further eye-tracking results combined these two areas to make one combined attention area, which compared against the full scene. Statistical significance tests were performed for all eye-tracking results and presented for Scheffe measures. Results showed that presentation summaries contained a much higher concentration of new information than full presentations. Users were found to be consistently more engaged for presentation summaries and spent a higher proportion of their time reading slides than for full presentations.

For additional information which can be extracted from eye-tracking procedures, combined heat maps from all tests were presented, with full gaze plots from all participants and all tests performed. Gaze plots in particular offer additional information as to levels of attention and focus typical for each video and version. Results showed that participants spent more time focused over the presentation slides during summaries than during full presentations. Gaze plots consistently showed shorter fixations during summaries which is consistent with higher cognitive activity.

Full results for comparisons against the use of an enhanced digital video browser were also presented, including questions asked of participants in addition to summed and averaged results. We also presented results from full comparisons with other versions of summaries, including those built using automatic classifications and keywords, automatic classifications using a subsection of available features in addition to keywords, and automatic classifications without using keywords. These results showed that use of these automatically generated presentation summaries compares favourably with the use of an enhanced digital video browser for gaining a quick, clear and concise overview of presentations, although these results were not statistically significant, limiting the conclusions which can be drawn from these.

In the next chapter we offer conclusions of the work presented in this thesis, providing answers to the research questions introduced in Chapter 1, and provide suggestions for future directions for this research.

Chapter 7

Conclusion and Future Work

In this chapter we review the contributions of this thesis. We provide answers to the research questions introduced in Chapter 1 and further discuss these. We conclude this chapter by providing suggestions of future directions for this research.

7.1 Summary of Thesis Contributions

This thesis focused on the classification of paralinguistic features automatically extracted from audio-visual recordings and academic presentations and on automatic summarisation of these presentations. We sought to summarise presentations by selected parts automatically classified as the most engaging, emphasised, and comprehensible for the audience.

We reported on the development of a multimodal corpus of academic talks that we annotated according to the quality of speakers and the estimated level of audience engagement. We studied the effects of acoustic and visual speech modalities on what human annotators perceived to be good speaking techniques. Using this dataset, we studied the correlation between these 'good' speaking techniques and direct audience engagement levels with the talk in progress, and also studied the correlation between annotated speaker ratings and actual audience engagement levels.

We trained a classifier to automatically predict a rating for the effectiveness of each speaker. Following this we showed how these speaker-based acoustic and visual modalities can be used for automatic prediction of audience engagement levels in a scenario where typical modalities associated with engagement detection such as eye-gaze are not feasible, by using visual information from the audience.

Previous work on emphasis detection in recordings of spoken content looked at the concept in the context of the audio-stream only (Arons, 1994; Kennedy and Ellis, 2003). In this thesis, we studied this phenomena using both audio and visual modalities, in the context of academic presentations. Our study shows that emphasis of speech depends very much upon speaker gesticulation in addition to audio pitch. However, speech intensity levels do not show any significant correlation with emphasis. These results demonstrate the importance of capturing gestures for detection of emphasis in audio-visual presentations.

We also report on the annotation of a multimodal corpus of academic presentations with the concept of audience comprehension. Using our labelled collection, we trained classifiers to predict potential comprehension levels for individual video segments. These were subsequently used to classify these concepts in presentation videos. We also investigated the use of early and late fusion strategies for the fusion of extracted audio-visual features, and their effects on performance in addition to calculating linear correlations between individual features and comprehension levels to investigate the effectiveness of multimodal features. This study showed that audio features and facial movement on the part of the speaker were found to influence audience comprehension. We demonstrated good results on classification of potential audience comprehension levels during academic presentations. The results show that the combination of information from multiple input streams affects comprehension levels, with audio features demonstrated to be the most important single modality, unsurprisingly, given that most of the information exists within the audio stream. This shows that it is possible to build a classifier to predict a speaker's potential to be comprehended. We also investigated the relationship between engagement and comprehension, and showed that it is possible to fully engage the audience to a presentation, whilst simultaneously failing to make the material fully comprehensible for the audience in question.

Following successful classification of the high level features extracted from the presentation, we developed an automatic video summarisation method to generate summaries of our academic presentations using a combination of the extracted high-level paralinguistic features, in combination with keywords taken from ASR speech transcripts.

The results of eye-tracking experiments performed on these summaries indicate that the generated summaries tend to contain a higher concentration of relevant information than full presentations, as indicated by the higher proportion of time participants spend carefully reading slides during summaries than during full presentations, and also by the lower proportion of time spent fixating on the speaker during summaries than during full presentations. We also found that watching generated presentation summaries is a good alternative to watching full presentations in which the viewer might not be fully interested. However, participants gave video summaries lower ratings than full video presentations for being easy to understand, enjoyable, informative and coherent.

More encouraging results were found for our comparisons between an enhanced digital video browser and the automatically generated summaries. This finding indicates that our generated summaries compare favourably with the use of an enhanced digital video browser, as currently used by searchers of video archives, though these results lack statistical significance. In this case generated video summaries received slightly higher scores than the enhanced digital video browser for 'ease of use' and effectiveness for 'gaining a clear and concise overview of the presentation'. However, they received slightly lower scores for 'tool of choice'. Given that the work of (Li et al., 2000) showed that an enhanced digital video browser scores very favourably for skimming presentation videos, and our own comparisons which rated our generated summaries at least equal in terms of effectiveness, this is a very promising result for these summarisation tasks.

The results of our study suggest that while classification of areas of engagement, emphasis and comprehension can be useful for summarisation, this may depend on how engaging and comprehensible videos are in the first place.

7.2 Research Questions Answered and Discussion

- *Research Question 1 (RQ-1): Can we build a classifier to automatically rate the qualities of a good public speaker?*

  In Chapter 4 of the thesis we showed that we can build a classifier to provide accurate ratings of the qualities of a public speaker. By training on a corpus of academic presentations annotated by human ratings for each presenter, we extracted audio-visual features from the video of each presentation. Linear correlations were calculated between individual audio-visual features and human annotation on speaker ratings to find the best performing features for classification tasks. We classified speaker ratings over an 8-class, 4-class and binary ranges, and showed how strong prediction accuracy on this task is achievable, despite the highly subjective nature of the task.

  Following the good accuracy shown for prediction of speaker ratings for the presenter to academic presentations, we looked to use these classifications to help us to address our second research question.

- *Research Question 2 (RQ-2): Can we build a classifier to automatically predict the levels of audience engagement by utilising speaker-based and basic visual audience-based modalities?*

  We also showed in Chapter 4 that by training a classifier on our corpus of academic presentations, annotated to estimated levels of audience engagement, we can predict levels of audience engagement to a good degree of accuracy. For the experiments performed in this thesis, we extracted a set of multimodal audio-visual features from the video of the presenter, in addition to video from the audience in order to extract motion information from the audience.

We again calculated linear correlations between individual multimodal features and human annotated audience engagement levels in order to discover the best performing modalities before classification tasks. We showed how classification results can be improved by fusing with classification outputs of speaker ratings from the previous research question. Results on this task are very promising and demonstrate the ability to predict levels of audience engagement based on the speaking techniques employed by the presenter.

Following classification of this concept, we calculated linear correlations between human annotations of speaker ratings and audience engagement levels in which we find a weak linear correlation between speaker ratings and audience engagement, which may be due to a larger than anticipated influence of the actual presentation content on end audience engagement levels.

We envisage that results could be improved further with the utilisation of more advanced feature extraction tools.

- *Research Question 3 (RQ-3-1): Can visual and acoustic stimuli on the part of the speaker be utilised to discover areas of special emphasis being provided by the speaker to indicate important parts of their presentation?*
  - Secondary RQ (RQ-3-2): *If we can detect spoken emphasis, is there a relationship between speaker ratings and emphasis, and between audience engagement and emphasis?*

We studied the identification of intentional or unintentional speaker emphasis in an audio-visual context. Our study showed that emphasis of speech in the audio-visual stream very much depends upon speaker gesticulation in addition to pitch. However, speech intensity levels did not show any significant correlation with emphasis. These results demonstrate the importance of gesturing for emphasis in the audio-visual stream. There was a lack of agreement between human annotators initially as they labelled emphasis points in video, which makes this a very difficult task for which to build a machine learning classifier.

We showed that it is possible to discover areas of special emphasis applied by the presenter, by finding areas of high pitch (top 20 percentile), in addition to areas of high gesticulation (top 20 percentile).

However, additional studies performed on correlations with emphasis and 'good' speaking techniques / engagement showed no real correlations between areas of 'good' public speaking techniques or with audience engagement levels. We also showed that high-levels of disagreement among human annotators for the annotation of this concept make it impractical to build a classifier to automatically predict speaker emphasis.

- *Research Question 4 (RQ-4-1): Can we utilise visual and acoustic stimuli to train a classifier to automatically identify levels of comprehension among the audience to a presentation?*

  - Secondary RQ (RQ-4-2): *If we can classify levels of audience comprehension, is there a relationship between audience engagement and audience comprehension?*

In Chapter 5 of the thesis we trained classifiers to predict potential comprehension levels for each video segment. We extracted a range of audio-visual features from the video of speaker and audience, in addition to visual features extracted from the presentation slides. We investigated the use of early and late fusion strategies for the fusion of extracted audio-visual features, and their effects on performance, in addition to calculating linear correlations between individual features and comprehension levels to investigate the effectiveness of individual multimodal features, in which audio features and facial movement were found to be most influential.

We achieved good results on classification of potential audience comprehension levels during academic presentations. Our results show that the combination of information from multiple input streams affects comprehension levels, with audio features unsurprisingly demonstrated to be the most important single modality, due to the majority of the information existing in the audio stream. We demonstrated that it is possible to build a classifier to predict a speakers potential to be comprehended.

We also investigated the relationship between engagement and comprehension and have found that it is possible to fully engage the audience to a presentation, whilst simultaneously failing to make the material fully comprehensible for the audience in question.

- *Research Question 5 (RQ-5): Can areas of special emphasis provided by the speaker, combined with detected areas of high audience engagement and high levels of audience comprehension, be used for effective summarisation of academic presentations?*

  In Chapter 6 of the thesis we performed a number of experiments in order to evaluate the effectiveness of summaries of academic presentations built using high-level paralinguistic features of ‘good’ speaking techniques, audience engagement, intentional or unintentional speaker emphasis and potential comprehension of the audience.

  We performed eye-tracking experiments in order to study a user’s levels of attention and focus to evaluate the effectiveness of full and summary presentations. We also evaluate these compared to the use of an enhanced digital video browser which has previously been shown to be very effective for gaining quick and clear overviews of presentations. Other comparisons were made with summaries built using a subset of the available features to evaluate further their effectiveness.

  Results of eye-tracking experiments indicate that the generated summaries tend to contain a higher concentration of relevant information than full presentations, as indicated by the higher proportion of time participants spend carefully reading slides during summaries than during full presentations, and also by the lower proportion of time spent fixating on the speaker during summaries than during full presentations. We find that watching generated presentation summaries is a good alternative to watching full presentations in which the viewer might not be fully interested. However the questionnaire used for this study did find that generated video summaries receive slightly lower ratings than full video presentations for being easy to understand, enjoyable, informative and coherent, however these findings were not statistically significant. Completed by participants upon completion of eye-tracking over the video, the purpose of the questionnaire was to gain additional data from participants in addition to their eye-tracking data about how they felt about presentations.

More encouraging results were shown by our comparisons with the enhanced digital video browser, which show that our generated summaries compare quite favourably to the use of an enhanced digital video browser, though these results lack statistical significance. In this case generated video summaries received slightly higher scores than the enhanced digital video browser for 'ease of use' and effectiveness for 'gaining a clear and concise overview of the presentation'. However, they received slightly lower scores for 'tool of choice'.

The results of our work indicate that while classification of areas of engagement, emphasis and comprehension can be useful for summarisation of academic presentations, this may depend on how engaging and comprehensible these videos are in the first place.

Results showed an increased fixation count for presentation summaries than for full presentations for all videos, confirming that users are indeed more attentive to presentation summaries. This difference is more pronounced for videos not already classified to be highly engaging, backing up the heat maps and questionnaire of eye-tracking participants which indicated that the summarisation process is more affective for videos which have not already been classified as highly engaging.

7.3 Potential Future Research Directions

In this section we propose some possible future directions for the research introduced in this thesis.

Possible further work includes the potential for further development of the work described in chapter 4 and 5 for the development of specific public speaking training systems. Research could be conducted on the alignment of virtual audience reactions to those of the real audience in this dataset. With a virtual audience giving realistic audience reactions to certain styles and speaking techniques of presenters, this paves the way for the development of more advanced public speaking training systems.

Further research extending that performed in Chapter 4 to examine correlations between additional individually extracted audio-visual features, combinations of such, and audience engagement levels would allow for the development of systems which can reliably inform a practising public speaker whether they should increase or decrease speech rate, pause rate, pitch or intensity, gesture more or less, move their head more or less, or make more eye-contact with the audience in order to engage the audience more in their presentations.

We speculate that improved accuracy over the existing paralinguistic features investigated in this research could be possible with improved feature extraction. One interesting research direction could be to track specific gestures on the part of the speaker and to study any correlation of these with audience engagement and with emphasis. Another interesting research direction would be the use of enhanced computer vision techniques to study facial expressions in the audience and investigate any perceived correlation of specific facial expressions coinciding with areas of high or low audience comprehension or engagement.

Further to this, specific gestures or facial expressions could also potentially be utilised to improve summarisation, to this end it is likely we would be looking for instances of a specific emotion prevalent on the presenter or even among the audience to such presentations. For example, the ability to summarise to instances where the speaker or audience are most happy, intrigued etc. could potentially provide more interesting ways of summarising presentations.

Another possible future direction for this research is the study of more enhanced video summarisation techniques. Increased use of natural language processing techniques could lead to the generation of more content dependant summaries. Further, the use of clustering techniques over lexical and low-level features of presentations such as low or high motion, pitch, intensity etc. could lead to interesting results.

With the availability of a larger data set than used for the work in this thesis, it would open the possibility of using deep learning to automatically find relevant features rather than using hand crafted features as used for the work in this thesis. Deep learning does not need hand crafted features provided as it is able to learn the features from the image data however as stated this requires a much larger data set than for hand crafted features as used in this work.

Laughter Detection is another possible direction for future research, this was outside of the scope of this thesis due to the complexity involved in detection of this feature, but future work detecting laughter in the audience could be of assistance, particularly for the detection of audience engagement.

7.4 Concluding Remarks

The work completed in this thesis summarised academic presentations by high-level paralinguistic features. We trained classifiers to predict a speaker rating for each presenter and to predict audience engagement levels during presentations. We developed a technique to automatically extract parts of intentional or unintentional emphasised speech. We have also trained a classifier to predict the speakers potential to be comprehended by their audience. A method was developed to automatically generate summaries of academic presentations by using the high-level features of speaker ratings, audience engagement, emphasised speech and audience comprehension, in addition to speech transcripts and keywords. Generated summaries were found to be more engaging, contain a higher concentration of new information than full presentations, and to aid users maintain focus for longer periods of time. We hope that this research will inspire further investigations into this topic, particularly into the summarisation of audio-visual material using high-level paralinguistic features.

Bibliography

Project 2: Local feature matching. https://www.cc.gatech.edu/~hays/compvision/results/proj2/html/delima3/index.html.

A tutorial on binary descriptors – part 3 – the orb descriptor. https://gilscvblog.com/2013/10/04/a-tutorial-on-binary-descriptors-part-3-the-orb-descriptor/, 2013.

Non verbally. https://appliedalliance.wordpress.com/2014/03/19/non-verbal-communication/, 2014.

Human and face detection from video (simulated streaming data). https://cloudmesh.github.io/introduction_to_cloud_computing/projects/nist_project_human_and_face_detection.html, 2015.

Optical character recognition (ocr). http://www.ni.com/example/30575/en/, 2015.

Loudness. https://www.sltinfo.com/loudness/, 2016. Accessed: 2016-07-11.

Opencv 3.1 tutorial optical flow (calcopticalflowfarneback). http://funvision.blogspot.ie/2016/02/opencv-31-tutorial-optical-flow.html, 2016.

Interest-operator. https://de.wikipedia.org/wiki/Interest-Operator, 2017.

Intensity - the physics hypertextbook. http://physics.info/intensity/, 2017. Accessed: 2017-06-18.

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

Y Alp Aslandogan and Clement T. Yu. Techniques and systems for image and video retrieval. *IEEE transactions on Knowledge and Data Engineering*, 11(1):56–63, 1999.

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F Smeaton, Georges Quénot, Maria Eskevich, Robin Aly, and Roeland Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVid*, volume 2016, 2016.

Ligia Batrinca, Giota Stratou, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer. Cicero-towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents*, pages 116–128. Springer, 2013.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

Roman Bednarik, Shahram Eivazi, and Michal Hradis. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*, page 10. ACM, 2012.

Dorothy Vera Margaret Bishop. Trog 2: Test for reception of grammar-version 2. *Edizioni Giunti OS, Firenze*, 2009.

Norman Blaikie. *Analyzing quantitative data: From description to explanation.* Sage, 2003.

Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345, 2002.

Dan Bohus and Eric Horvitz. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 2–9. ACM, 2014.

Francesca Bonin, Ronald Bock, and Nick Campbell. How do we react to context? annotation of individual and group engagement in a video corpus. In *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), and 2012 International Conference on Social Computing (SocialCom)*, pages 899–903. IEEE, 2012.

Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library.* O'Reilly Media, Inc., 2008.

Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39. Springer, 2005.

Modesto Castrillón-Santana, O Déniz-Suárez, L Antón-Canalís, and J Lorenzo-Navarro. Face and facial feature detection evaluation performance evaluation of public domain haar detectors for face and facial feature detection. pages 167–172, 2008.

Shih-Fu Chang and John R Smith. Extracting multidimensional signal features for content-based visual query. In *Visual Communications and Image Processing'95*, pages 995–1006. International Society for Optics and Photonics, 1995.

Gal Chechik, Eugene Ie, Martin Rehn, Samy Bengio, and Dick Lyon. Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 105–112. ACM, 2008.

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

Lei Chen, Gary Feng, Jilliam Joe, Chee Wee Leong, Christopher Kitchen, and Chong Min Lee. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 200–203. ACM, 2014.

S-F Cheng, William Chen, and Hari Sundaram. Semantic visual templates: linking visual features to semantics. In *Proceedings of 1998 International Conference on Image Processing. ICIP98*, pages 531–535. IEEE, 1998.

Chinese University of Hong Kong. The chinese university of hong kong department of obstetrics and gynaecology. `http://department.obg.cuhk.edu.hk/researchsupport/IntraClass_correlation.asp`, 2016. Accessed: 2016-04-22.

Yu-Jen Chou, Bor-Chyun Wang, and Shun-Feng Su. Enhance intelligence video surveillance with depth and color information. In *2013 International Conference on System Science and Engineering (ICSSE)*, pages 19–24. IEEE, 2013.

Lee J Corrigan, Christina Basedow, Dennis Küster, Arvid Kappas, Christopher Peters, and Ginevra Castellano. Mixing implicit and explicit probes: finding a ground truth for engagement in social human-robot interactions. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pages 140–141. ACM, 2014.

Keith Curtis, Gareth J.F. Jones, and Nick Campbell. Effects of good speaking techniques on audience engagement. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pages 35–42. ACM, 2015.

Keith Curtis, Gareth J.F. Jones, and Nick Campbell. Speaker impact on audience comprehension for academic presentations. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 129–136. ACM, 2016.

Keith Curtis, Gareth J.F. Jones, and Nick Campbell. Identification of emphasised regions in audio-visual presentations. In *Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016), Copenhagen, 29-30 September 2016*, number 141, pages 37–42. Linköping University Electronic Press, 2017a.

Keith Curtis, Gareth J.F. Jones, and Nick Campbell. Utilising high-level features in summarisation of academic presentations. In *Proceedings of the 2017 ACM International Conference on Multimedia Retrieval*, pages 315–321. ACM, 2017b.

Keith Curtis, Nick Campbell, and Gareth J.F. Jones. Development of an annotated multimodal dataset for the investigation of classification and summarisation of presentations using high-level paralinguistic features. In *LREC*, 2018a.

Keith Curtis, Gareth J.F. Jones, and Nick Campbell. Summarising academic presentations using linguistic and paralinguistic features. In *Proceedings of the 2nd International Conference on Human Computer Interaction Theory and Applications (HUCAPP)*, 2018b.

Nivja H De Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, 2009.

Matthijs Douze, Dan Oneata, Mattis Paulin, Clément Leray, Nicolas Chesneau, Danila Potapov, Jakob Verbeek, Karteek Alahari, Cordelia Schmid, Lori Lamel, et al. The inria-lim-vocr and axes submissions to trecvid 2014 multimedia event detection. *Proceedings of TRECVID 2014, NIST, USA*, 2014.

Bruno Dumas, Denis Lalanne, and Sharon Oviatt. Multimodal interfaces: A survey of principles, models and frameworks. *Human machine interaction*, pages 3–26, 2009.

Marta Dynel. Turning speaker meaning on its head: non-verbal communication and intended meanings. *Pragmatics & Cognition*, 19(3):422–447, 2011.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 835–838. ACM, 2013.

Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *12th European Conference on Machine Learning*, pages 145–156. Springer, 2001.

Eibe Frank, Mark Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H Witten, and Len Trigg. Weka-a machine learning workbench for data mining. In *Data Mining and Knowledge Discovery Handbook*, pages 1269–1277. Springer, 2010.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37. Springer, 1995.

Andrew C Gallagher and Tsuhan Chen. Understanding images of groups of people. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 256–263. IEEE, 2009.

Joanna Garner and Michael Alley. How the design of presentation slides affects audience comprehension: A case for the assertion–evidence approach. *International Journal of Engineering Education*, 29(6):1564–1579, 2013.

Daniel Gatica-Perez, L McCowan, Dong Zhang, and Samy Bengio. Detecting group interest-level in meetings. In *ICASSP-05: 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–489. IEEE, 2005.

Raymond W Gibbs. *Intentions in the Experience of Meaning.* Cambridge University Press, 1999.

Joseph F Grafsgaard, Joseph B Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 42–49. ACM, 2014.

H Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1991.

Magnus Haake, Kristina Hansson, Agneta Gulz, Susanne Schötz, and Birgitta Sahlén. The slower the better? does the speaker's speech rate influence children's performance on a language comprehension test? *International Journal of Speech-Language Pathology*, 16(2):181–190, 2014.

G Hall. Pearson's correlation coefficient. *Other Words*, 1(9), 2015.

Alan Hanjalic and HongJiang Zhang. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(8):1280–1289, 1999.

Zongbo Hao, Qianni Zhang, Ebroul Ezquierdo, and Nan Sang. Human action recognition by fast dense trajectories. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 377–380. ACM, 2013.

Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, volume 15, page 50, 1988.

A Hauptmann and M Smith. Text, speech, and vision for video segmentation: The informedia project. In *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision*, 1995.

Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. Auto-summarization of audio-video presentations. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, pages 489–498. ACM, 1999.

Geoffrey Holmes, Andrew Donkin, and Ian H Witten. Weka: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361. IEEE, 1994.

Mark J Huiskes, Bart Thomee, and Michael S Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 527–536. ACM, 2010.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73. Association for Computational Linguistics, 2010.

Nakamasa Inoue, Zhuolin Liang, Mengxi Lin, Tran Hai Dang, Koichi Shinoda, Zhang Xuefeng, and Kazuya Ueki. Tokyotech-waseda at trecvid 2014. In *Proceedings of TRECVID workshop*, pages 1–13, 2014.

Toastmasters International. *Competent Communication A Practical Guide to Becoming a Better Speaker*, 2015.

Susan Jamieson et al. Likert scales: how to (ab) use them. *Medical education*, 38 (12):1217–1218, 2004.

Minsu Jang, Cheonshu Park, Hyun-Seung Yang, Jae-Hong Kim, Young-Jo Cho, Dong-Wook Lee, Hye-Kyung Cho, Young Kim, Kyoungwha Chae, and Byeong-Kyu Ahn. Building an automated engagement recognizer based on video analysis. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pages 182–183. ACM, 2014.

Hideo Joho, Joemon M Jose, Roberto Valenti, and Nicu Sebe. Exploiting facial expressions for affective video summarisation. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 31. ACM, 2009.

Shanon X Ju, Michael J Black, Scott Minneman, and Don Kimber. Summarization of videotaped presentations: automatic analysis of motion and gesture. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):686–696, 1998.

Adam Kendon. *Gesture: Visible action as utterance.* Cambridge University Press, 2004.

Lyndon S Kennedy and Daniel PW Ellis. Pitch-based emphasis detection for characterization of meeting recordings. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*, pages 243–248. IEEE, 2003.

Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2): 155–163, 2016.

Robert E Kraut and Robert E Johnston. Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology*, 37(9): 1539, 1979.

Shiri Lev-Ari. Comprehending non-native speakers: theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, 5, 2014.

Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2 (1):1–19, 2006.

Congcong Li, Yi-Ta Wu, Shiaw-Shian Yu, and Tsuhan Chen. Motion-focusing key frame extraction and video summarization for lane surveillance system. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 4329–4332. IEEE, 2009.

Francis C Li, Anoop Gupta, Elizabeth Sanocki, Li-wei He, and Yong Rui. Browsing digital video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 169–176. ACM, 2000.

Ying Li, Tong Zhang, and Daniel Tretter. An overview of video abstraction techniques. Technical report, Technical Report HPL-2001-191, HP Laboratory, 2001.

Rainer W Lienhart. Dynamic video summarization of home video. In *Electronic Imaging*, pages 378–389. International Society for Optics and Photonics, 1999.

Jackson Liscombe, Jennifer Venditti, and Julia Hirschberg. Classifying subject ratings of emotional speech using acoustic features. In *Eighth European Conference on Speech Communication and Technology*, 2003.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Xiaofei Lu. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496, 2010.

Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, volume 81, pages 674–679, 1981.

Michael Lutter. Mel-frequency cepstral coefficients. `http://recognize-speech.com/feature-extraction/mfcc`, 2015. Accessed: 2015-07-01.

Marcin Marszalek and Cordelia Schmid. Semantic hierarchies for visual object recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07.*, pages 1–7. IEEE, 2007.

MAURO ANDREA VOICE STUDIO - CANTO AULAS. Mauro andrea voice studio - canto aulas. `http://www.estudiodevoz.com.br/2015/07/formantes-ressonancias-sintonizacao-de_28.html`, 2015. Accessed: 2016-06-11.

Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (3):305–317, 2005.

David McNeill. *Hand and mind: What gestures reveal about thought.* University of Chicago Press, 1992.

David McNeill. *Language and gesture*, volume 2. Cambridge University Press, 2000.

Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 116:374–388, 1976.

Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke. The meeting project at icsi. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–7. Association for Computational Linguistics, 2001.

Desmond Morris and Desmond Morris. Manwatching: A field guide to human behavior. Technical report, 1977.

Jeho Nam and Ahmed H Tewfik. Video abstract of video. In *1999 IEEE 3rd Workshop on Multimedia Signal Processing*, pages 117–122. IEEE, 1999.

Azra Nasreen and G Shobha. Key frame extraction from videos-a survey. *International Journal of Computer Science & Communication Networks*, 3(3):194, 2013.

Carlton W Niblack, Ron Barber, Will Equitz, Myron D Flickner, Eduardo H Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos, and Gabriel Taubin. Qbic project: querying images by content, using color, texture, and shape. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pages 173–187. International Society for Optics and Photonics, 1993.

Catharine Oertel and Giampiero Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 99–106. ACM, 2013.

Catharine Oertel, Céline De Looze, Stefan Scherer, Andreas Windmann, Petra Wagner, and Nick Campbell. Towards the automatic detection of involvement in conversation. In *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, pages 163–170. Springer, 2011.

Harry Ferdinand Olson. *Music, physics and engineering*, volume 1769. Courier Corporation, 1967.

Paul Over, Jon Fiscus, Greg Sanders, David Joy, Martial Michel, George Awad, Alan Smeaton, Wessel Kraaij, and Georges Quénot. Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVid*, page 52, 2014.

Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *Sixth International Conference on Computer Vision, 1998.*, pages 555–562. IEEE, 1998.

Nikesh J Patel and Manali S Rajput. A review on different keyframe abstraction techniques from the video. *International Journal of Advance Research in Computer Science and Management Studies*, 2014.

Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. Video digests: a browsable, skimmable format for informational lecture videos. In *ACM Symposium on User Interface Software and Technology*, pages 573–582, 2014.

Lluís Payrató. Non-verbal communication. *Key Notions for Pragmatics. Amsterdam/Philadelphia: John Benjamins*, pages 163–194, 2009.

Jiang Peng and Qin Xiao-Lin. Keyframe-based video summary using visual attention clues. *IEEE MultiMedia*, (2):64–73, 2009.

Yuxin Peng, Jian Zhang, Panpan Tang, Lei Huang, Xin Huang, and Xiangteng He. Pku-icst at trecvid 2014: Instance search task. In *Proceedings of TRECVID workshop*, pages 1–6, 2014.

Keith Rayner and Sara C Sereno. Eye movements in reading: Psycholinguistic studies. *Handbook of Psycholinguistics*, pages 57–81, 1994.

Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006.

Andrew Rosenberg. Autobi-a tool for automatic tobi annotation. In *INTERSPEECH*, pages 146–149, 2010.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571. IEEE, 2011.

Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

Joerg Schulenburg. Gocr. *Available on: http://www-e. unimagdeburg. de/jschulen/ocr*, 2010.

Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315. Citeseer, 2009.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330. ACM Press, 2006.

Michael A Smith and Takeo Kanade. *Video skimming for quick browsing based on audio and image characterization.* School of Computer Science, Carnegie Mellon University, 1995.

Michael A Smith and Takeo Kanade. Video skimming and characterization through the combination of image and language understanding. In *1998 IEEE International Workshop on Content-Based Access of Image and Video Database*, pages 61–70. IEEE, 1998.

Spoken Data Video Processing. Spoken data. `http://spokendata.com/`, 2016. Accessed: 2015-07-08.

Eva Strangert. What makes a good speaker? subjective ratings and acoustic measurements. In *Proceedings from Fonetik 2007: speech, music and hearing, quarterly progress and status report, TMH-QPSR, Vol 50, 2007*, pages 29–32, 2007.

Eva Strangert and Joakim Gustafson. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *INTERSPEECH*, volume 8, pages 1688–1691, 2008.

Johan Sundberg et al. *The acoustics of the singing voice.* Scientific American, 1977.

Super Lectures Video Hosting. Super lectures. `http://superlectures.com/`, 2016. Accessed: 2015-07-08.

Martin Szummer and Rosalind W Picard. Indoor-outdoor image classification. In *1998 IEEE International Workshop on Content-Based Access of Image and Video Database, 1998.*, pages 42–51. IEEE, 1998.

Annie H Takeuchi and Stewart H Hulse. Absolute pitch. *Psychological bulletin*, 113 (2):345, 1993.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634, 1995.

Hai Tao and Thomas S Huang. Connected vibrations: a modal analysis approach for non-rigid motion tracking. In *1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998.*, pages 735–740. IEEE, 1998.

Candemir Toklu, S-P Liou, and Madirakshi Das. Videoabstract: a hybrid approach to generate semantically meaningful video summaries. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1333–1336. IEEE, 2000.

Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1):37, 2007.

Abhishek Verma, Suket Arora, and Preeti Verma. Ocr-optical character recognition. In *7th International Conference on Recent Innovations in Science, Engineering and Management*, 2016.

Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492. IEEE, 2010.

Chengde Zhang, Xiao Wu, Mei-Ling Shyu, and Qiang Peng. Adaptive association rule mining for web video event classification. In *2013 IEEE 14th International Conference on Information Reuse and Integration (IRI)*, pages 618–625. IEEE, 2013.

Rong Zhao and William I Grosky. Narrowing the semantic gap-improved text-based web document retrieval using visual features. *IEEE Transactions on Multimedia*, 4(2):189–200, 2002.

Glossary of Term Definitions

Engagement For the definition of engagement as referred to in this thesis we refer to the Oxford English Dictionary, who define engagement as ***The action of engaging or being engaged***, with the synonyms of ***participation***, ***participating***, ***taking part***, ***sharing***, ***partaking***, ***involvement***, and ***association***.

Eye Gaze (The direction of) a person's gaze; frequently attributive, especially designating various technologies based on detecting the point on which a person's eyes are focused.

Visual Attention The procedure by which one object, the objective, is chosen for study from among many competitor objects, the distractors. At its most basic, attention is defined as the process by which we select a subset from all of the available information for further processing.

Non-Verbal Not involving or using words or speech. In the context of this thesis the term non-verbal communication refers to all non-spoken elements of communication.

Paralinguistic In the context of this thesis, the term paralinguistic refers to the high-level concepts of audience engagement, audience comprehension, and emphasised speech.

Appendix A

A.1 Publications

The research presented in this dissertation appeared in several peer-reviewed conference proceedings. The work on engagment detection, presented in chapter 4 was published in (Curtis et al., 2015). The work on emphasis detection, also presented in chapter 4 was published in (Curtis et al., 2017a). The work on audience comprehension, presented in chapter 5, was published in (Curtis et al., 2016). Finally, the summarisation aspect of this work, presented in chapter 6, is published in the papers (Curtis et al., 2017b) and (Curtis et al., 2018b). Finally, details of the collection and annotation of the multimodal dataset used in this research and presented in chapter 3 of this thesis are published in (Curtis et al., 2018a).

- Keith Curtis, Gareth J F Jones and Nick Campbell. "Effects of Good Speaking Techniques on Audience Engagement", 17th ACM International Conference on Multimodal Interaction, November 2015.

- Shu Chen, Keith Curtis, David N Racca, Liting Zhou, Gareth J F Jones and Noel E O'Connor. "DCU ADAPT @ TRECVid 2015: Video Hyperlinking Task", TRECVid Workshop, November 2015.

- Keith Curtis, Gareth J F Jones and Nick Campbell. "Identification of Emphasised Regions in Audio-Visual Presentations", The 4th European and 7th Nordic Symposium on Multimodal Communication, September 2016.

- Keith Curtis, Gareth J F Jones and Nick Campbell. "Speaker Impact on Audience Comprehension for Academic Presentations", 18th ACM International Conference on Multimodal Interaction, November 2016.

- Keith Curtis, Gareth J F Jones and Nick Campbell. "Utilising High-Level Features in Summarisation of Academic Presentations", 7th ACM International Conference on Multimedia Retrieval, June 2017.

- Keith Curtis, Gareth J F Jones and Nick Campbell. “Summarising Academic Presentations using Linguistic and Paralinguistic Features”, 2nd International Conference on Human Computer Interaction Theory and Applications, January 2018.

- Keith Curtis, Nick Campbell and Gareth J F Jones. “Development of an Annotated Multimodal Dataset for the Investigation of Classification and Summarisation of Presentations using High-Level Paralinguistic Features”, LREC, May 2018.

Appendix B

B.1 Full Eye-Tracking Results

This Appendix contains full results of eye-tracking experiments performed in Chapter 6. Results for both LSD and Scheffe measures of statistical significance are listed. This Appendix lists results from eye-tracking evaluation not deemed significant enough for inclusion in Chapter 6.

Table B.1: Eye-tracking Video 1 - slides scene compared by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.02	0.05	0.655
full	summ	FD.M	LSD	0.02	0.05	0.655
summ	full	FD.M	Scheffe	-0.02	0.05	0.655
full	summ	FD.M	Scheffe	0.02	0.05	0.655
summ	**full**	**percent**	**LSD**	**11.07**	**4.97**	**0.043**
full	**summ**	**percent**	**LSD**	**-11.07**	**4.97**	**0.043**
summ	**full**	**percent**	**Scheffe**	**11.07**	**4.97**	**0.043**
full	**summ**	**percent**	**Scheffe**	**-11.07**	**4.97**	**0.043**
summ	**full**	**FCp100s**	**LSD**	**37.32**	**14.32**	**0.021**
full	**summ**	**FCp100s**	**LSD**	**37.32**	**14.32**	**0.021**
summ	**full**	**FCp100s**	**Scheffe**	**37.32**	**14.32**	**0.021**
full	**summ**	**FCp100s**	**Scheffe**	**37.32**	**14.32**	**0.021**

Table B.2: Eye-tracking Video 1 - speaker scene compared by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.29	0.22	0.213
full	summ	FD.M	LSD	0.29	0.22	0.213
summ	full	FD.M	Scheffe	-0.29	0.22	0.213
full	summ	FD.M	Scheffe	0.29	0.22	0.213
summ	full	percent	LSD	-8.51	5.10	0.117
full	summ	percent	LSD	8.51	5.10	0.117
summ	full	percent	Scheffe	-8.51	5.10	0.117
full	summ	percent	Scheffe	8.51	5.10	0.117
summ	full	FCp100s	LSD	-2.07	6.79	0.765
full	summ	FCp100s	LSD	2.07	6.79	0.765
summ	full	FCp100s	Scheffe	-2.07	6.79	0.765
full	summ	FCp100s	Scheffe	2.07	6.79	0.765

Results for video 1, Table B.1 and Table B.2, show a statistically significant difference between the summary and full versions, for the percentage of time fixated per scene and the number of fixations per 100 seconds. As demonstrated, and more clearly visible in the core figures in Table 6.4, this indicates a significant difference in the amount of time users spent fixating on the slides while watching the summary than the full video. This indicates that users found there to be a much higher concentration of new information during the summary than during the slides.

Table B.3: Eye-tracking Video 2 - slides scene compared by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.08	0.08	0.337
full	summ	FD.M	LSD	0.08	0.08	0.337
summ	full	FD.M	Scheffe	-0.08	0.08	0.337
full	summ	FD.M	Scheffe	0.08	0.08	0.337
summ	full	percent	LSD	-2.72	7.90	0.735
full	summ	percent	LSD	2.72	7.90	0.735
summ	full	percent	Scheffe	-2.72	7.90	0.735
full	summ	percent	Scheffe	2.72	7.90	0.735
summ	full	FCp100s	LSD	7.04	11.16	0.538
full	summ	FCp100s	LSD	-7.04	11.16	0.538
summ	full	FCp100s	Scheffe	7.04	11.16	0.538
full	summ	FCp100s	Scheffe	-7.04	11.16	0.538

Table B.4: Eye-tracking Video 2 - speaker scene compared by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	0.17	0.22	0.471
full	summ	FD.M	LSD	-0.17	0.22	0.471
summ	full	FD.M	Scheffe	0.17	0.22	0.471
full	summ	FD.M	Scheffe	-0.17	0.22	0.471
summ	full	percent	LSD	7.27	7.45	0.346
full	summ	percent	LSD	-7.27	7.45	0.346
summ	full	percent	Scheffe	7.27	7.45	0.346
full	summ	percent	Scheffe	-7.27	7.45	0.346
summ	full	FCp100s	LSD	1.55	6.81	0.823
full	summ	FCp100s	LSD	-1.55	6.81	0.823
summ	full	FCp100s	Scheffe	1.55	6.81	0.823
full	summ	FCp100s	Scheffe	-1.55	6.81	0.823

While again in results for video 2, Table B.3 and Table B.4, key differences can be observed in the average fixation duration per scene, and to a lesser extent in the fixation count per 100 seconds, more clearly visible from the core figures in Table 6.4, neither of these differences are statistically significant.

Table B.5: Eye-tracking Video 3 - slides scene compared by version

I	J	Variable	Measure	Diff	Error	Sig
summ	**full**	**FD.M**	**LSD**	**-0.15**	**0.05**	**0.009**
full	**summ**	**FD.M**	**LSD**	**0.15**	**0.05**	**0.009**
summ	**full**	**FD.M**	**Scheffe**	**-0.15**	**0.05**	**0.009**
full	**summ**	**FD.M**	**Scheffe**	**0.15**	**0.05**	**0.009**
summ	full	percent	LSD	-2.67	6.54	0.689
full	summ	percent	LSD	2.67	6.54	0.689
summ	full	percent	Scheffe	-2.67	6.54	0.689
full	summ	percent	Scheffe	2.67	6.54	0.689
summ	full	FCp100s	LSD	16.91	13.54	0.232
full	summ	FCp100s	LSD	-16.91	13.54	0.232
summ	full	FCp100s	Scheffe	16.91	13.54	0.232
full	summ	FCp100s	Scheffe	-16.91	13.54	0.232

Table B.6: Eye-tracking Video 3 - speaker scene compared by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	0.01	0.29	0.976
full	summ	FD.M	LSD	-0.01	0.29	0.976
summ	full	FD.M	Scheffe	0.01	0.29	0.976
full	summ	FD.M	Scheffe	-0.01	0.29	0.976
summ	full	percent	LSD	4.98	6.62	0.464
full	summ	percent	LSD	-4.98	6.62	0.464
summ	full	percent	Scheffe	4.98	6.62	0.464
full	summ	percent	Scheffe	-4.98	6.62	0.464
summ	full	FCp100s	LSD	-2.13	7.55	0.782
full	summ	FCp100s	LSD	2.13	7.55	0.782
summ	full	FCp100s	Scheffe	-2.13	7.55	0.782
full	summ	FCp100s	Scheffe	2.13	7.55	0.782

Results for video 3, Table B.5 and Table B.6, show that there is a significant difference in the mean fixation length over the slides during the summary than during the full presentation. As can be seen by looking at the core figures in Table 6.4, The averaged fixation duration is much shorter for fixations on the slides. This indicates that users tend to look at the slides in order to follow and absorb information printed in the slides.

Table B.7: Eye-tracking Video 4 - slides scene compared by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.06	0.05	0.266
full	summ	FD.M	LSD	0.06	0.05	0.266
summ	full	FD.M	Scheffe	-0.06	0.05	0.266
full	summ	FD.M	Scheffe	0.06	0.05	0.266
summ	**full**	**percent**	**LSD**	**15.68**	**6.42**	**0.028**
full	**summ**	**percent**	**LSD**	**-15.68**	**6.42**	**0.028**
summ	**full**	**percent**	**Scheffe**	**15.68**	**6.42**	**0.028**
full	**summ**	**percent**	**Scheffe**	**-15.68**	**6.42**	**0.028**
summ	**full**	**FCp100s**	**LSD**	**54.96**	**16.98**	**0.006**
full	**summ**	**FCp100s**	**LSD**	**-54.96**	**16.98**	**0.006**
summ	**full**	**FCp100s**	**Scheffe**	**54.96**	**16.98**	**0.006**
full	**summ**	**FCp100s**	**Scheffe**	**-54.96**	**16.98**	**0.006**

Table B.8: Eye-tracking video 4 - speaker scene compared by version

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.20	0.14	0.161
full	summ	FD.M	LSD	0.20	0.14	0.161
summ	full	FD.M	Scheffe	-0.20	0.14	0.161
full	summ	FD.M	Scheffe	0.20	0.14	0.161
summ	full	percent	LSD	-6.40	4.43	0.170
full	summ	percent	LSD	6.40	4.43	0.170
summ	full	percent	Scheffe	-6.40	4.43	0.170
full	summ	percent	Scheffe	6.40	4.43	0.170
summ	full	FCp100s	LSD	-2.10	5.47	0.707
full	summ	FCp100s	LSD	2.10	5.47	0.707
summ	full	FCp100s	Scheffe	-2.10	5.47	0.707
full	summ	FCp100s	Scheffe	2.10	5.47	0.707

Results for video 4, Table B.7 and Table B.8, show a statistically significant difference between the summary and full versions, for the percentage of time fixated per scene and the number of fixations per 100 seconds. As demonstrated in the core figures in Table 6.4, this indicates a significant difference in the amount of time users spent fixating on the slides while watching the summary than the full video. This indicates that users found there to be a much higher concentration of new information during the summary than during the slides.

B.1.1 Further Evaluations of Gaze

By combining the slides and speaker scene's from the last set of eye-tracking evaluations into one combined attention scene, and comparing with the whole scene, we can evaluate further differences between full presentations and summaries in how much they both entice full attention to the presentations in progress.

Table B.9: Eye-tracking Video 1 - full scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.09	0.06	0.163
full	summ	FD.M	LSD	0.09	0.06	0.163
summ	full	FD.M	Scheffe	-0.09	0.06	0.163
full	summ	FD.M	Scheffe	0.09	0.06	0.163
summ	full	percent	LSD	2.36	1.68	0.181
full	summ	percent	LSD	-2.36	1.68	0.181
summ	full	percent	Scheffe	2.36	1.68	0.181
full	summ	percent	Scheffe	-2.36	1.68	0.181
summ	**full**	**FCp100**	**LSD**	**36.73**	**17.12**	**0.050**
full	**summ**	**FCp100**	**LSD**	**-36.73**	**17.12**	**0.050**
summ	**full**	**FCp100**	**Scheffe**	**36.73**	**17.12**	**0.050**
full	**summ**	**FCp100**	**Scheffe**	**-36.73**	**17.12**	**0.050**

Table B.10: Eye-tracking Video 1 - attention scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.09	0.06	0.159
full	summ	FD.M	LSD	0.09	0.06	0.159
summ	full	FD.M	Scheffe	-0.09	0.06	0.159
full	summ	FD.M	Scheffe	0.09	0.06	0.159
summ	full	percent	LSD	2.57	2.05	0.232
full	summ	percent	LSD	-2.57	2.05	0.232
summ	full	percent	Scheffe	2.57	2.05	0.232
full	summ	percent	Scheffe	-2.57	2.05	0.232
summ	**full**	**FCp100**	**LSD**	**35.57**	**16.13**	**0.045**
full	**summ**	**FCp100**	**LSD**	**-35.57**	**16.13**	**0.045**
summ	**full**	**FCp100**	**Scheffe**	**35.57**	**16.13**	**0.045**
full	**summ**	**FCp100**	**Scheffe**	**-35.57**	**16.13**	**0.045**

Tables B.9 and B.10 show a statistically significant difference between the summary and full versions, for the number of fixations per 100 seconds. Taking these results in conjunction with the overall core figures in Table 6.9 show that video 1 summary is more engaging than the overall presentation video for video 1.

Table B.11: Eye-tracking Video 2 - full scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.01	0.08	0.865
full	summ	FD.M	LSD	0.01	0.08	0.865
summ	full	FD.M	Scheffe	-0.01	0.08	0.865
full	summ	FD.M	Scheffe	0.01	0.08	0.865
summ	**full**	**percent**	**LSD**	**5.68**	**2.41**	**0.033**
full	**summ**	**percent**	**LSD**	**-5.68**	**2.41**	**0.033**
summ	**full**	**percent**	**Scheffe**	**5.68**	**2.41**	**0.033**
full	**summ**	**percent**	**Scheffe**	**-5.68**	**2.41**	**0.033**
summ	full	FCp100	LSD	9.67	14.51	0.516
full	summ	FCp100	LSD	-9.67	14.51	0.516
summ	full	FCp100	Scheffe	7.04	14.51	0.516
full	summ	FCp100	Scheffe	-7.04	14.51	0.516

Table B.12: Eye-tracking Video 2 - attention scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.01	0.08	0.862
full	summ	FD.M	LSD	0.01	0.08	0.862
summ	full	FD.M	Scheffe	-0.01	0.08	0.862
full	summ	FD.M	Scheffe	0.01	0.08	0.862
summ	full	percent	LSD	4.55	2.89	0.138
full	summ	percent	LSD	-4.55	2.89	0.138
summ	full	percent	Scheffe	4.55	2.89	0.138
full	summ	percent	Scheffe	-4.55	2.89	0.138
summ	full	FCp100	LSD	8.31	11.59	0.485
full	summ	FCp100	LSD	-8.31	11.59	0.485
summ	full	FCp100	Scheffe	8.31	11.59	0.485
full	summ	FCp100	Scheffe	-8.31	11.59	0.485

Again in video 2 results, Table B.11 and Table B.12, key differences are observed in the average fixation duration per scene, and to a lesser extent in the fixation count per 100 seconds, more clearly visible from the figures in Table 6.9, neither of these differences are statistically significant. Participants still spend a higher proportion of their time fixating on the attention scene for video summaries than for full video presentations.

Table B.13: Eye-tracking Video 3 - full scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.02	0.09	0.813
full	summ	FD.M	LSD	0.02	0.09	0.813
summ	full	FD.M	Scheffe	-0.02	0.09	0.813
full	summ	FD.M	Scheffe	0.02	0.09	0.813
summ	full	percent	LSD	7.53	3.63	0.057
full	summ	percent	LSD	-7.53	3.63	0.057
summ	full	percent	Scheffe	7.53	3.63	0.057
full	summ	percent	Scheffe	-7.53	3.63	0.057
summ	full	FCper100	LSD	11.22	15.24	0.474
full	summ	FCper100	LSD	-11.22	15.24	0.474
summ	full	FCper100	Scheffe	11.22	15.24	0.474
full	summ	FCper100	Scheffe	-11.22	15.24	0.474

Table B.14: Eye-tracking Video 3 - attention scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.11	0.11	0.333
full	summ	FD.M	LSD	0.11	0.11	0.333
summ	full	FD.M	Scheffe	-0.11	0.11	0.333
full	summ	FD.M	Scheffe	0.11	0.11	0.333
summ	full	percent	LSD	2.31	3.24	0.488
full	summ	percent	LSD	-2.31	3.24	0.488
summ	full	percent	Scheffe	2.31	3.24	0.488
full	summ	percent	Scheffe	-2.31	3.24	0.488
summ	full	FCper100	LSD	14.78	15.23	0.348
full	summ	FCper100	LSD	-14.78	15.23	0.348
summ	full	FCper100	Scheffe	14.78	15.23	0.348
full	summ	FCper100	Scheffe	-14.78	15.23	0.348

Tables B.13 and B.14 show that there is no statistically significant difference between the two scene's of the video, however, there is a large, but not significant difference in the percentage of time spent fixating on the attention scene during the video summary compared with during the full video presentation.

Table B.15: Eye-tracking Video 4 - full scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.11	0.06	0.080
full	summ	FD.M	LSD	0.11	0.06	0.080
summ	full	FD.M	Scheffe	-0.11	0.06	0.080
full	summ	FD.M	Scheffe	0.11	0.06	0.080
summ	full	percent	LSD	0.72	3.62	0.846
full	summ	percent	LSD	-0.72	3.62	0.846
summ	full	percent	Scheffe	0.72	3.62	0.846
full	summ	percent	Scheffe	-0.72	3.62	0.846
summ	**full**	**FCp100**	**LSD**	**45.01**	**15.27**	**0.011**
full	**summ**	**FCp100**	**LSD**	**-45.01**	**15.27**	**0.011**
summ	**full**	**FCp100**	**Scheffe**	**45.01**	**15.27**	**0.011**
full	**summ**	**FCp100**	**Scheffe**	**-45.01**	**15.27**	**0.011**

Table B.16: Eye-tracking Video 4 attention scene compared by **version**

I	J	Variable	Measure	Diff	Error	Sig
summ	full	FD.M	LSD	-0.10	0.05	0.099
full	summ	FD.M	LSD	0.10	0.05	0.099
summ	full	FD.M	Scheffe	-0.10	0.05	0.099
full	summ	FD.M	Scheffe	0.10	0.05	0.099
summ	**full**	**percent**	**LSD**	**9.28**	**3.84**	**0.030**
full	**summ**	**percent**	**LSD**	**-9.28**	**3.84**	**0.030**
summ	**full**	**percent**	**Scheffe**	**9.28**	**3.84**	**0.030**
full	**summ**	**percent**	**Scheffe**	**-9.28**	**3.84**	**0.030**
summ	**full**	**FCp100**	**LSD**	**52.86**	**16.14**	**0.006**
full	**summ**	**FCp100**	**LSD**	**-52.86**	**16.14**	**0.006**
summ	**full**	**FCp100**	**Scheffe**	**52.86**	**16.14**	**0.006**
full	**summ**	**FCp100**	**Scheffe**	**-52.86**	**16.14**	**0.006**

Tables B.15 and B.16 demonstrate a statistically significant difference between the summary and full versions, for the percentage of time fixated per scene and the number of fixations per 100 seconds. This indicates that users found there to be a much higher concentration of new information during the summary than the full version.

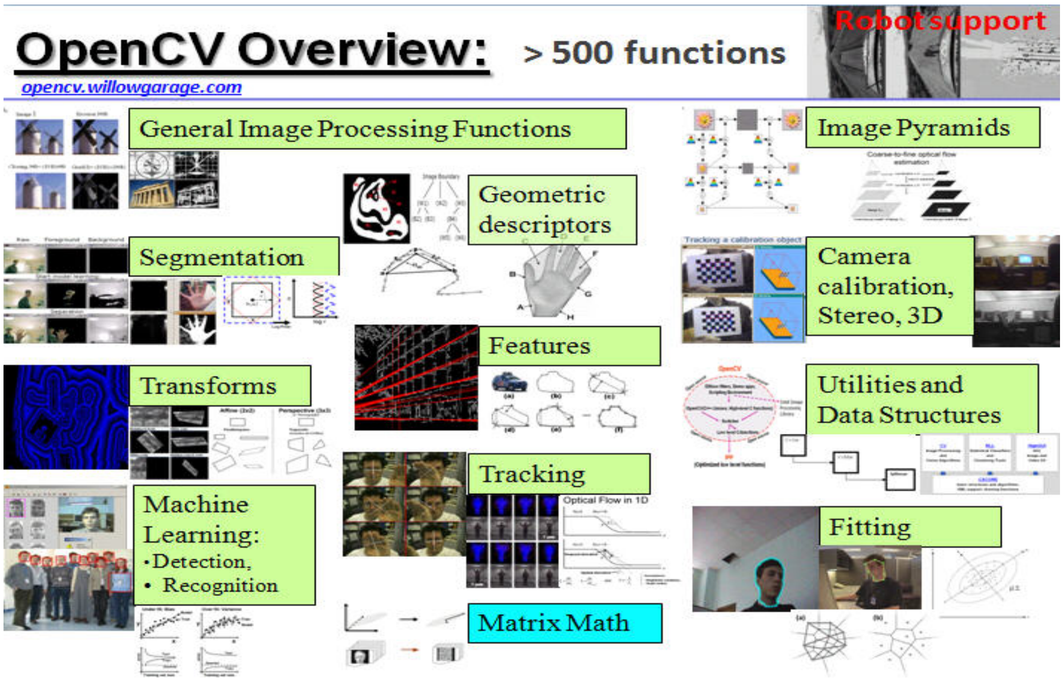
Appendix C

This chapter contains technical information on some of the tools used within this thesis. Feature Extraction for this research was performed using OpenCV (Bradski and Kaehler, 2008), OpenSMILE (Eyben et al., 2013), and Praat (Boersma et al., 2002). Further, Machine Learning algorithms and techniques were trained and experimented upon using Weka data mining workbench (Frank et al., 2010).

C.1 Technical Information

C.1.1 OpenCV

OpenCV (Open Source Computer Vision) is a library of programming functions mainly aimed at real-time computer vision (Bradski and Kaehler, 2008). Its library is cross-functional and free for use under the open source BSD license. It is written in C++ and its primary interface is C++, though it still retains a less comprehensive though extensive C interface. There are bindings in Python, Java and Matlab. Its runs on a variety of platforms including Windows, Linux, OS X, Android and Blackberry. Its application areas include 2D and 3D feature toolkits, facial recognition, motion tracking, object identification, segmentation and recognition, structure from motion and stereopsis stereo vision.



OpenCV Overview: > 500 functions

opencv.willowgarage.com

Robot support

General Image Processing Functions

Geometric descriptors

Image Pyramids

Segmentation

Camera calibration, Stereo, 3D

Features

Transforms

Utilities and Data Structures

Tracking

Machine Learning:
- Detection,
- Recognition

Fitting

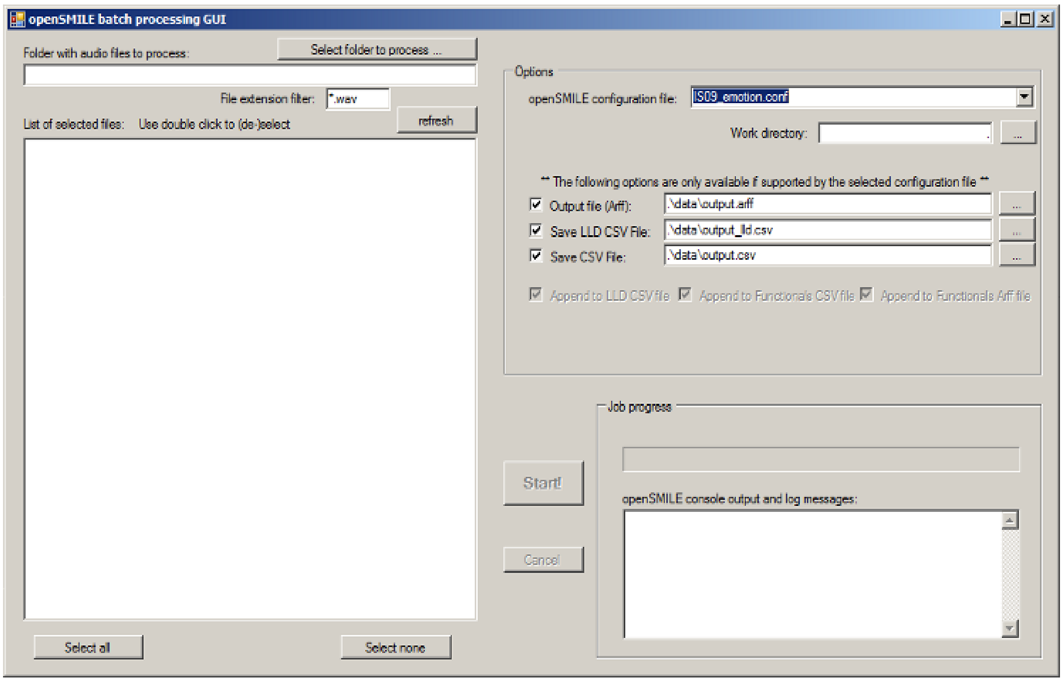
Matrix Math

Optical Flow in 1D
Tracking a calibration object
Coarse-to-fine optical flow estimation
OpenCV

Figure C.1: OpenCV functions

C.1.2 OpenSMILE

Speech & Music Interpretation by Large-space Extraction: openSMILE feature extraction tool allows for extraction of large audio feature spaces in real-time, combining features from Speech Processing and Music Information Retrieval. It is written in C++ and is available both as a dynamic library and as a standalone command-line executable. Feature extraction components can be freely interconnected to create new and customfeatures, and new components can be added via an easy binary plugin interface.



openSMILE batch processing GUI

File extension filter: *.wav

openSMILE configuration file: IS09_emotion.conf

Work directory: .

** The following options are only available if supported by the selected configuration file **

☑ Output file (Arff): .\data\output.arff

☑ Save LLD CSV File: .\data\output_lld.csv

☑ Save CSV File: .\data\output.csv

☑ Append to LLD CSV file ☑ Append to Functionals CSV file ☑ Append to Functionals Arff file

Folder with audio files to process:
Select folder to process ...
List of selected files: Use double click to (de-)select
refresh
Options
...
Job progress
Start!
Cancel
openSMILE console output and log messages:
Select all
Select none

Figure C.2: OpenSMILE GUI

C.1.3 Praat

Praat (Boersma et al., 2002) is a free scientific software package for the analysis of speech in phonetics. It runs on a wide range of operating systems including Windows, Unix, Linux and Mac. Praat is useful for spectral analysis, pitch analysis, formants analysis and intensity analysis. It is useful for annotating sound objects and files and is also useful for manipulation of pitch, intensity, duration and formants. A range of scripts are publicly available for a wide variety of tasks which can be accomplished using Praat.

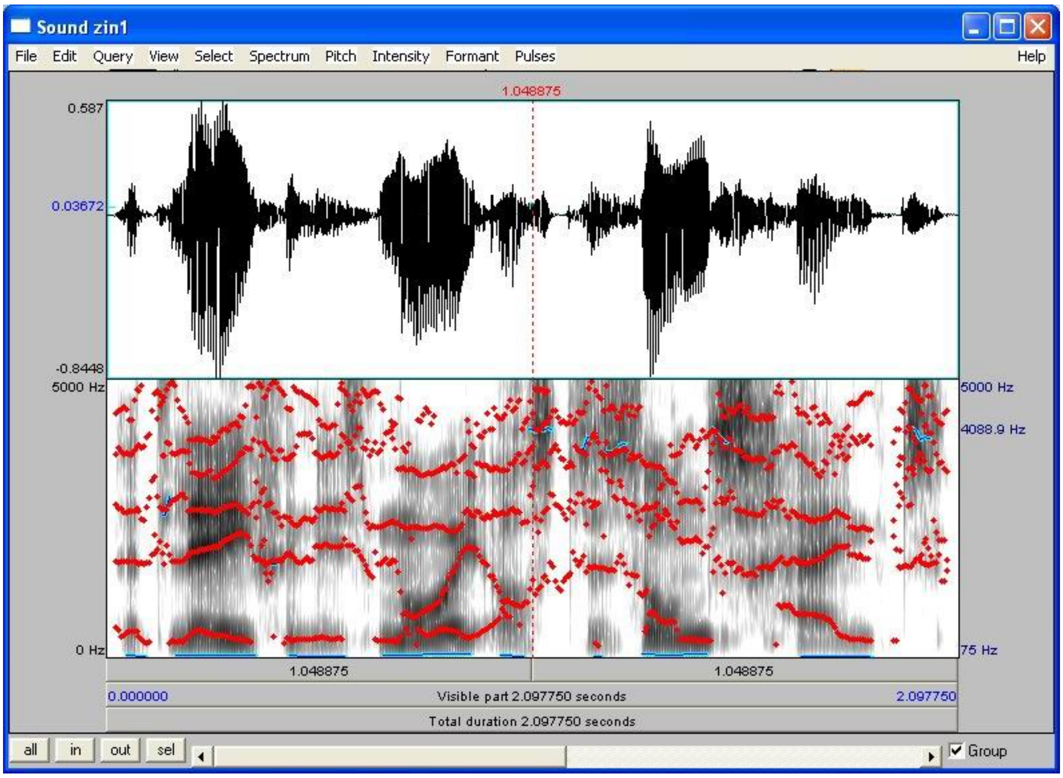
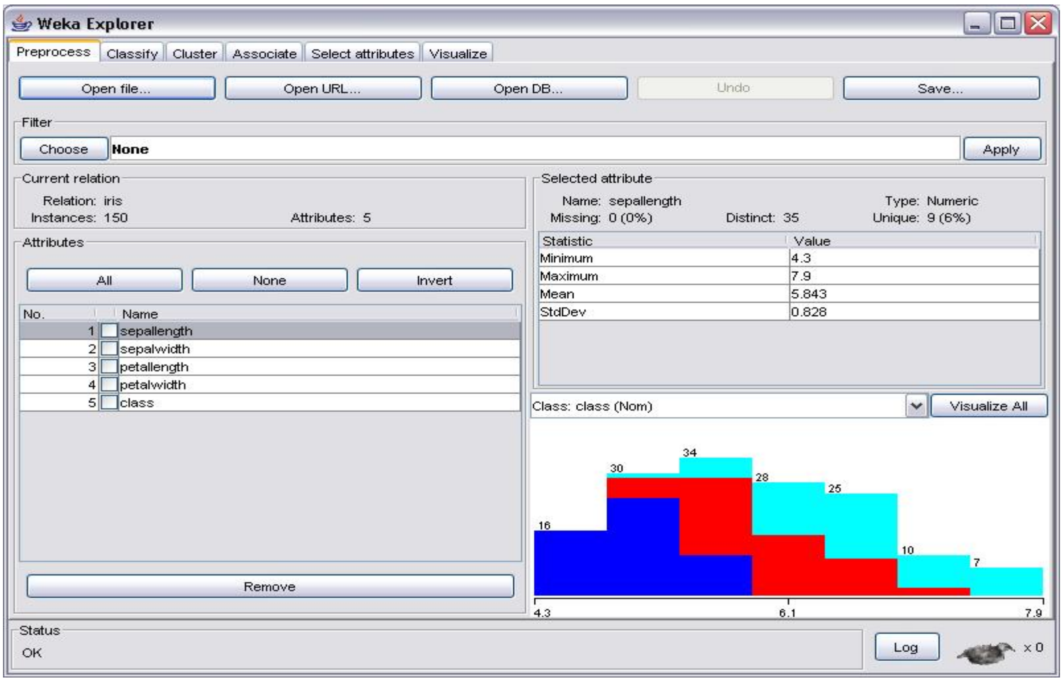


Figure C.3: Praat Screenshot. The Oscillogram (upper panel) and spectrogram (lower panel) of a sentence.

C.1.4 Weka

Waikato Environment for Knowledge Analysis (Weka) (Holmes et al., 1994) is a suite of Machine Learning algorithms for Data Mining tasks. Weka is written in the Java programming language and developed at the University of Waikato, New Zealand. Weka contains a collection of visualisation tools and algorithms for data analysis and predictive modelling. Weka is freely available under the GNU General Public License, is freely portable and contains a comprehensive collection of data processing and modelling techniques. Weka supports data modelling, pre-processing, clustering, regression, visualisation and feature selection. It's main user interface is the Explorer, but it's functionality can also be accessed through the command line. The original version of Weka was developed in 1993 and was a mix of Tcl/Tk, C and makefiles. The decision was taken in 1997 to completely re-develop Weka from scratch using Java.



Filter

None

Current relation

Relation: iris

Instances: 150 Attributes: 5

Attributes

No.		Name
1	□	sepallength
2	□	sepalwidth
3	□	petallength
4	□	petalwidth
5	□	class

Selected attribute

Name: sepallength Type: Numeric

Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)

Status

OK

Figure C.4: Weka Explorer