Summarising Academic Presentations using Linguistic and Paralinguistic Features

Keith Curtis[1], Gareth J. F. Jones[1] and Nick Campbell[2]
[1]*ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland*
[2]*ADAPT Centre, School of Computer Science & Statistics, Trinity College Dublin, Dublin, Ireland*

Keywords: Video Summarisation, Classification, Evaluation, Eye-Tracking.

Abstract: We present a novel method for the automatic generation of video summaries of academic presentations using linguistic and paralinguistic features. Our investigation is based on a corpus of academic conference presentations. Summaries are first generated based on keywords taken from transcripts created using automatic speech recognition (ASR). We augment spoken phrases by incorporating scores for audience engagement, comprehension and speaker emphasis. We evaluate the effectiveness of our summaries generated for individual presentations by performing eye-tracking evaluation of participants as they watch summaries and full presentations, and by questionnaire of participants upon completion of eye-tracking studies. We find that automatically generated summaries tend to maintain the user's focus and attention for longer, with users losing focus much less often than for full presentations.

1 INTRODUCTION

Online archives of multimedia content are growing rapidly. Every minute in 2016, 300 hours of new video material was uploaded to YouTube, while 2.78 million videos were viewed. It is time consuming and a growing challenge for users to be able to browse content of interest in such multimedia archives - either in response to user queries or in informal exploration of content. The goal of this work is to provide an effective and efficient way to summarise audio-visual recordings where the significant information is primarily in the audio stream.

Existing multimedia information retrieval research has focused on matching text queries against written meta-data or transcribed audio (Chechik et al., 2008), in addition to seeking to match visual queries to low-level features such as colours, textures, shape and object recognition, plus scene type classification - urban, countryside, or places etc. (Huiskes et al., 2010). In the case of multimedia content where the information is primarily visual, content can be represented by multiple keyframes extracted using methods such as object recognition and facial detection. These are matched against visual queries, and retrieved videos shown to the user using keyframe surrogates. Matching of the visual component of these queries is generally complemented by text search against a transcript of any available spoken audio and any meta-data provided (Lew et al., 2006).

The above methods are limited to content with significant visual dimension or where spoken content makes the subject clear or relevance is to some extent unambiguous. However, significant amounts of multimedia content does not have these features, e.g. public presentations such as lectures 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. In this research, we explore novel methods of summarising academic presentations, where most of the information exists in the audio stream, using linguistic and paralinguistic features, and evaluate the effectiveness of these automatically generated summaries.

In this study we hypothesise that the classification of areas of audience engagement, speaker emphasis, and the speaker's potential to be comprehended by the audience, can be used to improve summarisation methods for academic presentations. We address the following research question "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 audio-visual recordings of presentations?" We evaluate this summarisation approach using eye-tracking and by questionnaire.

Curtis, K., Jones, G. and Campbell, N.
Summarising Academic Presentations using Linguistic and Paralinguistic Features.
DOI: 10.5220/0006720500640073
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 2: HUCAPP*, pages 64-73
ISBN: 978-989-758-288-2

This paper is structured as follows: Section 2 introduces related work in video summarisation in addition to describing the classification of high level features. Section 3 introduces the multimodal corpus used for our experiments, while Section 4 describes the procedure for creating automatic summaries. This is followed by Section 5 which describes the evaluation tasks performed and their results. Finally, conclusions are offered in Section 6 of the paper.

2 PREVIOUS WORK

This section looks at related work on summarisation and skimming of audio-visual recording of academic presentations.

(Ju et al., 1998) present a system for analysing and annotating video sequences of technical talks. They use a robust motion estimation technique to detect key frames and segment the video into subsequences containing a single slide. Potential gestures are tracked using active contours. This first automatic video analysis system helps users to access presentation videos intelligently.

(He et al., 1999)s use prosodic information from the audio stream to identify speaker emphasis during presentations, in addition to pause information. They develop three summarisation algorithms focusing on: slide transition based summarisation, pitch activity based summarisation and summarisation based on slide, pitch and user-access information. In their work they found that speaker emphasis did not provide sufficient information to generate effective summaries.

(Joho et al., 2009) captures and analyses 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. Results suggest that there are at least two or three distinguished parts of videos that can be seen as the highlight by various viewers.

Work in (Pavel et al., 2014) created a set of tools for creating video digests of informational video. Informal evaluation suggests that these tools make it easier for authors of informational talks to create video digests. They found that video digests afford browsing and skimming better than alternative video presentation techniques.

We aim to extend these by classifying the most engaging and comprehensible parts of presentations and identifying emphasised regions within them before summarising them, to include highly rated regions of such high-level concepts, in addition to keywords taken from the transcript of the presentation.

2.1 High-Level Concept Classification

The novel video summarisation method reported in this study incorporates the high level concepts of audience engagement, emphasised speech and the speakers potential to be comprehended. In this section we overview previous work on the development of these high-level concept detectors.

2.1.1 Classification of Audience Engagement

Prediction of audience engagement levels and applying ratings for 'good' speaking techniques was performed by (Curtis et al., 2015). This was achieved by first employing human annotators to watch video segments of presentations and to provide a rating of how good that speaker was at presenting the material. Audience engagement levels were measured in a similar manner by having annotators watch video segments of the audience to academic presentations, and providing estimates of just how engaged the audience appeared to be as a whole. Classifiers were trained on extracted audio-visual features using an Ordinal Class Classifier.

It was demonstrated that the qualities of a 'good' speaker can be predicted to an accuracy of 73% over a 4-class scale. Using speaker-based techniques alone, audience engagement levels can be predicted to an accuracy of 68% over the same scale. By combining with basic visual features from the audience as whole, this can be improved to 70% accuracy.

2.1.2 Identification of Emphasised Speech

Identification of intentional or unintentional emphasised speech was performed by (Curtis et al., 2017a). This was achieved by having human annotators label areas of emphasised speech. Annotators were asked to watch through short video clips and to mark areas where they considered the speech to be emphasised, either intentionally or unintentionally. Basic audio-visual features of audio pitch and visual motion were extracted from the data. From the analysis performed, it was clear that speaker emphasis occurred during areas of high pitch, but also during areas of high visual motion coinciding with areas of high pitch.

Candidate emphasised regions were marked from extracted areas of pitch within the top 1, 5, and top 20 percentile of pitch values, in addition to top 20 percentile of gesticulation down to the top 40 percentile of values respectively. All annotated areas of emphasis contained 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 was not accompanied by pitch in the top 20 percentile.

2.1.3 Predicting the Speakers Potential to be Comprehended

Prediction of audience comprehension was performed by (Curtis et al., 2016). In this work, human annotators were recruited through the use of crowd-sourcing. Annotators were asked to watch each section of a presentation and to first provide a textual summary of the contents of that section of the presentation and following this to provide an estimate of how much they comprehended the material during that section. Audio-visual features were extracted from video of the presenter in addition to visual features extracted from video of the audience, and OCR over the slides for each presentation. Additional fluency features were also extracted from the speaker audio. Using the above described extracted features, a classifier was trained to predict the speaker's potential to be comprehended.

It was demonstrated that it is possible to build a classifier to predict potential audience comprehension levels, obtaining accuracy over a 7-class range of 52.9%, and over a binary classification problem to 85.4%.

3 MULTIMODAL CORPUS

No standard publicly available dataset exists for work of this nature. Since we require recordings of the audience and of the speaker for academic presentations, for this work, we used the International Speech Conference Multi-modal corpus (ISCM) from (Curtis et al., 2015). This contains 31 academic presentations totalling 520 minutes of video, with high quality 1080p parallel video recordings of both the speaker and the audience to each presentation. We chose four videos from this dataset for our evaluation of our video summaries to ensure good coverage but to avoid too much evaluations. Analysis of these four videos showed them to be: the most engaging, the least engaging, the most comprehensible videos, and the video with highest presentation rankings, and they were selected use in our summarisation study for this reason.

4 CREATION OF PRESENTATION SUMMARIES

This section describes the steps involved in the generation of presentation summaries. Summaries were generated using ASR transcripts, significant keywords extracted form transcripts, and annotated values for 'good' public speaking techniques, audience engagement, intentional or unintentional speaker emphasis and the speakers potential to be comprehended. Human annotated values of these paralinguistic features were used for summarisation for eye-tracking experiments to ensure these summaries used the best possible classifications of these features. Numbered sections here can be related to numbers appearing in algorithm 1. Presentation summaries are generated to between 20% and 30% of original presentation lengths for this evaluation.

1. Using the ASR transcripts, we use pause information, which gives start and end times for each spoken phrase during the presentation. This provides a basis for the separation of presentations at the phrase-level.
2. We first apply a ranking for each phrase based on the number of significant keywords extracted form transcripts, or words of significance, contained within it. For the first set of baseline summaries, we generate summaries by using the highest ranking sentences.
   Following this, the additional ranking is applied to each sentence based on human annotations of 'good' speaking techniques.
3. Speaker Ratings 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.
4. Following this, audience engagement annotations are also applied directly. We take the true annotated engagement level and apply this ranking to each sentence contained within each segment throughout the presentation.
5. As emphasis was annotated for all videos, we use automatic classifications for intentional or unintentional speaker emphasis. For each classification of emphasis, we apply an additional ranking 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 sentence

within that segment. For weightings, we choose to half the Speaker Rating annotation, while choosing to keep the original for other annotations, this is to reduce the impact of Speaker Ratings on the final output. Points of emphasis receive a ranking of 1, while keywords receive a ranking of 2, in order to prioritise the role of keywords in the summary generation process.

To generate the final set of video summaries, the highest ranking 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 apply a minimum size to our generated video summaries. Final selected segments are then joined together to generate small, medium and large summaries for each presentation.

Algorithm 1: Generate Summaries.

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

5 EVALUATION OF VIDEO SUMMARIES

Presentation summaries are only as effective as they have been found to be by their target audience. In this paper we provide a comprehensive evaluation of generated presentation summaries in order to discover the effectiveness of this summarisation strategy. In this regard, we carry out our study using an eye-tracking system in which participants watch full presentations and separate presentation summaries. From studying the eye-movements and focus of participants we can make inferences as to how engaged participants were as they watched presentations and summaries.

Eye-tracking 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 allows us to understand clearly whether generated summaries have any effect on levels of attention / engagement of participants as they watch presentation summaries.

Questionnaires were also provided to participants in order to discover how useful and effective the participants considered the presentation summaries to be. Also, by summarising using only a subset of all available features, we aimed to discover how effective the individual features are by crowdsourcing a separate questionnaire on presentation summaries generated using subsets of available features. Features used for this further evaluation were: full feature classifications, visual only classifications, audio only classifications, and full feature classifications with no keywords.

5.1 Gaze-Detection Evaluation

For the eye-tracking, participants watched one full presentation whilst having their eye-movements tracked. Participants also watched a separate presentation summary, again whilst having their eye-movements tracked. The question being addressed here was whether or not participants retained attention for longer periods to the presentations for summaries than for full presentations, to test the hypothesis that summaries were engaging and comprehensible.

The eye-tracking study was completed by a total of 24 separate participants. As there were 4 videos to be evaluated in total, eight different test condition were developed, with 4 participants per test. This allowed for full variation of the order in which participants watched the videos. Therefore, half of all participants began by watching a full presentation and finished by watching a summary of a separate presentation. The other half began by watching a presentation summary and finished by watching a full, separate presentation. This was to prevent any issues of bias or fatigue from influencing these results.

Table 1 shows the core values for eye-tracking measurements per video, version and scene. The 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, where version 1 corresponds to the video summary, and version 2 to the full video. The overall scene is 1 and the attention scene - the area around the slides and the speaker is 2. Measurements obtained include: number of fixations, mean length of fixations, total sum of fixation lengths, percentage of time fixated per scene, fixation count and number of fixations per second.

Again, from Table 1, 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, where this figure is consistently higher for summaries than for full presentations. Again, this is evidence of increased levels of participant 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, 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 of participants for summaries than for full presentations.

Table 2 shows a statistically significant ($p < 0.05$) difference between summary and full versions of video 1, for the number of fixations per 100 seconds. These results indicate that video 1 summary is more engaging than the full presentation for video 1. For video 2, statistically significant differences ($p < 0.05$) are observed in the average fixation duration per scene, and to a lesser, not statistically significant, extent in the fixation count per 100 seconds. Participants still spend a higher proportion of their time fixating on the attention scene for summaries than for full presentations.

Video 3 results show a large difference between the two scene's of the video, there is a statistically significant ($p < 0.1$) difference in the percentage of time spent fixating on the attention scene during the summary compared with full presentation. Video 4 shows a statistically significant difference between the summary and full versions, for the number of fixations per 100 seconds ($p < 0.05$). This video also shows a statistically significant difference ($p < 0.1$) for the mean fixation duration between full presentation and the presentation summary. This indicates that users found there to be a much higher concentration of new information during the summary than the full version. These differences can be inspected further by looking back to Table 1.

5.2 Gaze Plots

In this section, we show gaze plots from our eye-tracking study. Gaze plots are data visualisations which can communicate important aspects of visual behaviour clearly. By looking carefully at plots for full and summary videos, the difference in attention and focus for different video types becomes more clearly defined. For each video, 4 representative gaze plots are chosen, 2 on top from full presentations, and 2 below from summaries.

From the representative gaze plots in Figure 1 we can see that participants hold much higher levels of attention during summaries than for full presentations, with far less instances of them losing focus or looking around the scene, instead focussing entirely on the slides and speaker. The many small circles over the slides area represent a large number of smaller fixations - indicating high engagement.

From the representative plots in Figure 2 we see large improvements in summaries over the full presentations. While participants still lose focus on occasion, and improvements from full presentations is not as refined as for the previous video, large improvements are still gained, with the vast majority of fixations taking place over the presentation slides and the speaker. For comparison, gaze plots for the full video shows that fixations tended to be quite dispersed.

From the representative plots in Figure 3 we see how the number of occasions on which participants lose focus is reduced, with a big improvement on full presentations. Gaze plots show the difference for this video much better than the statistical tests in the previous section do. For full presentations, fixations are very dispersed with large numbers of fixations away from the slides and speakers. Summaries show a large improvement with a much reduced number of instances of participants losing focus.

From the representative plots in Figure 4 we can see that while summaries are imperfect, with instances of participants losing attention, huge improvements in attention and focus are made, although this may depend on how engaging the videos were in the first place. While summaries for Video 4 (speechRT6) still show some instances of participants losing focus, the original full presentation was found to be the least engaging video of the dataset. This is also noticeable from gaze plots. The gaze plots show a high number of fixations away form the slides and presenters. Gaze plots of summaries also show smaller fixations than full presentation gaze plots, which indicates higher levels of engagement for presentation summaries, in addition to the obvious position of these fixations taking place predominantly over the slides and speakers.

Table 1: Totals per video, version, scene.

Vid	Version	Scene	Fixations	mean fixation	time fixated	% fixated	fixations	F.P.S.
1	Summ	Whole	432.5	0.492	203.986	94.438	432.625	2.00289
1	Summ	Atten	417.37	0.495	199.17	92.208	417.375	1.93229
1	Full	Whole	1580	0.582	889.486	92.079	1580	1.63561
1	Full	Atten	1523	0.587	865.952	89.643	1523	1.5766
2	Summ	Whole	311.87	0.695	209.284	95.129	311.875	1.41761
2	Summ	Atten	293.25	0.71	201.996	91.816	293.25	1.33295
2	Full	Whole	1153.12	0.709	780.864	89.446	1153.125	1.32087
2	Full	Atten	1091.12	0.724	761.826	87.265	1091.125	1.24987
3	Summ	Whole	224.37	0.62	135.407	91.491	224.375	1.51605
3	Summ	Atten	193.87	0.694	130.091	87.899	193.875	1.30997
3	Full	Whole	1076.75	0.641	643.995	83.963	1076.75	1.40385
3	Full	Atten	891.37	0.8	656.5	85.593	891.375	1.16216
4	Summ	Whole	406.37	0.431	169.388	89.624	406.375	2.15013
4	Summ	Atten	385.5	0.466	174.105	92.119	385.5	2.03968
4	Full	Whole	1591.25	0.536	832.177	88.908	1591.25	1.70005
4	Full	Atten	1414.37	0.561	775.37	82.839	1414.375	1.51108

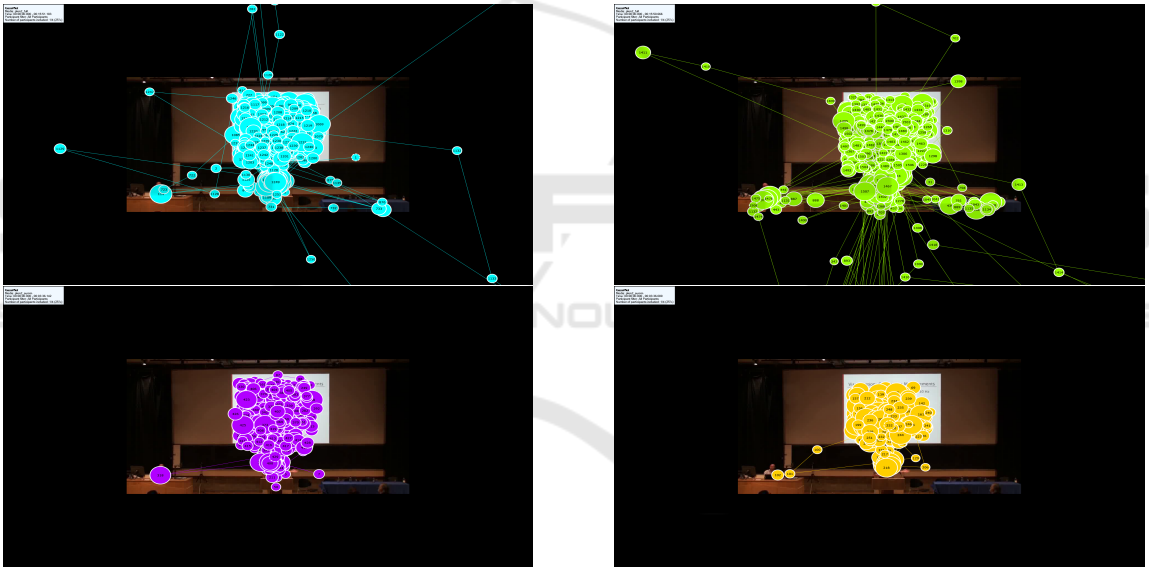

Figure 1: Plen 2 - Representative Gaze Plots.

Overall, gaze plots show many fewer instances of participants losing focus from the presentation. Gaze during summaries is primarily focussed on the presentation slides as users gain more new information, with deviations from this usually reverting back to the speaker. Also visible from gaze plots of Video 1 (plen2) and particularly Video 4 (speechRT6), is the shorter fixations (smaller circles) for summaries than for full presentations. This can be seen more clearly by looking back to the figures reported in Table 1.

5.3 Questionnaire Evaluation of Summary Types

In the next part of our summary evaluation, we illicit questionnaire responses from participants who watched summaries generated using all available features or just a subset of available features. Different summaries were generated using all available features, audio features only, visual features only, or audio-visual features with no keywords used. From this we aimed to discover the importance of different features on presentation summaries. A total of 48 participants watched the summaries and answered the

Table 2: Eye-tracking videos scene by **version**.

I	J	Variable	Measure	Diff	Error	Sig (p value)
Video 1						
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**
Video 2						
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
Video 3						
summ	full	FD.M	Scheffe	-0.02	0.09	0.813
summ	**full**	**percent**	**Scheffe**	**7.53**	**3.63**	**0.057**
summ	full	FCp100	Scheffe	11.22	15.24	0.474
Video 4						
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**

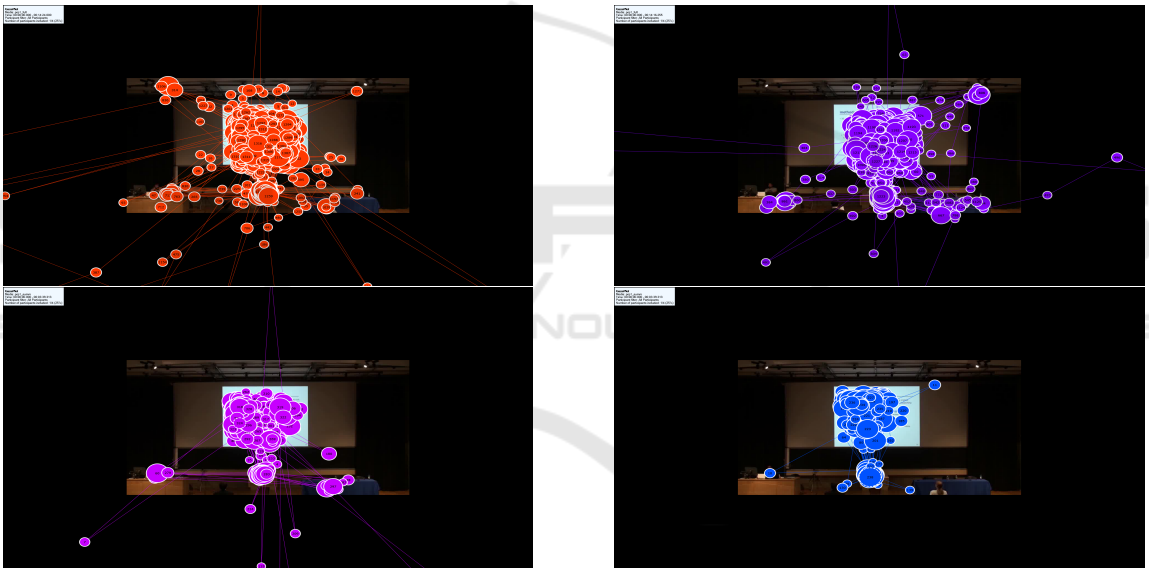

Figure 2: prp 1 - Representative Gaze Plots.

questionnaire on each summary. Each video was evaluated by 12 participants in total. The order in which participants watched summaries was also alternated to avoid issues of bias in these results.

In Table 3 we show further evaluations between summaries built using all available features, and summaries built using just 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 by using only audio features, with visual features not being considered. Similarly, for visual only summaries, classification of these features was performed using only visual features, with audio features not being considered. For no keyword summaries, classification of these features is performed and the only information excluded were keywords. For Classify summaries, all available information is used for generating summaries. For this actual classification outputs are used rather than ground truth human annotations for most engaging, comprehensible parts of presentations. Results in this table reflect Likert-scale rankings of participants level of agreement with each statement.

1. This summary is easy to understand.
2. This summary is informative.
3. This summary is enjoyable.

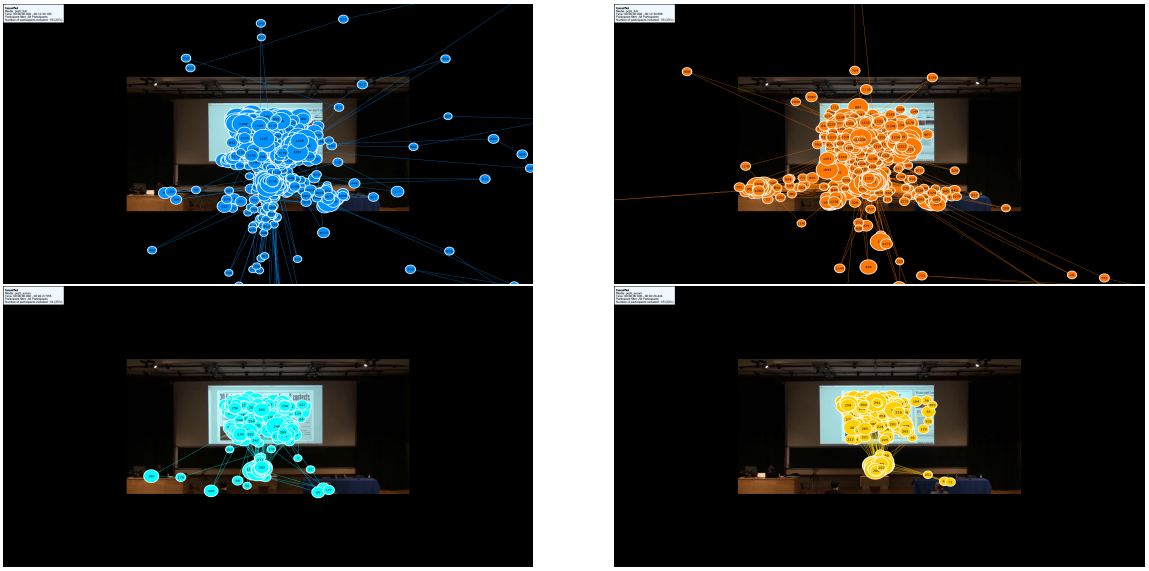

Figure 3: prp 5 - Representative Gaze Plots.

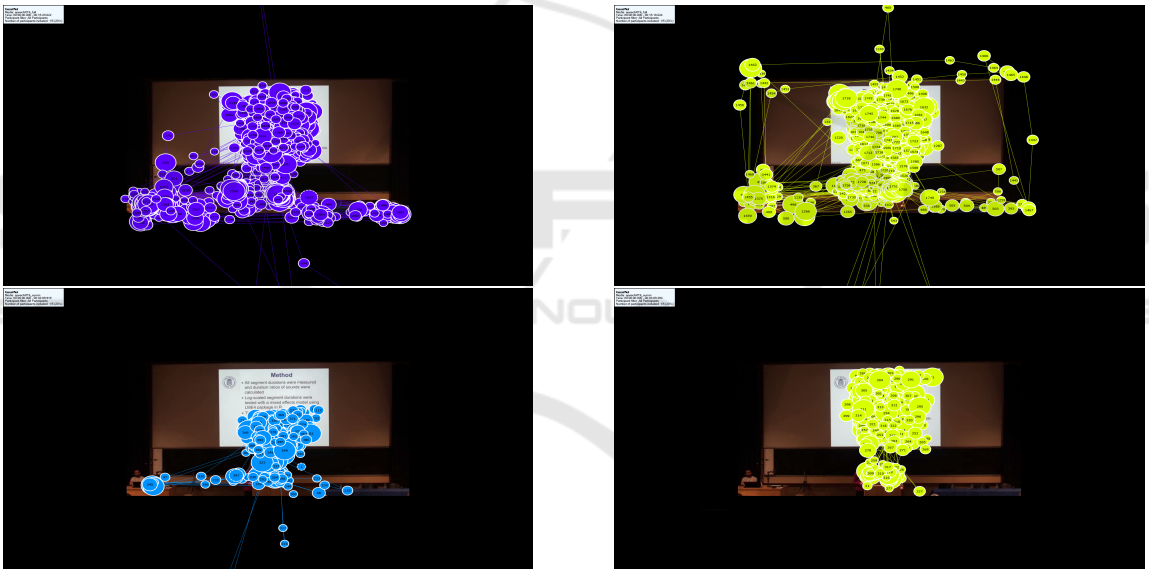

Figure 4: speechRT6 - Representative Gaze Plots.

4. This summary is coherent.
5. This summary would aid me in deciding whether to watch the full video.

From Table 3, we can see that the results of audio-only classifications and visual-only classifications result in summaries which are rated less easy to understand and informative than summaries built using full information and with no key words. Summaries built using no keywords also lack coherence, while summaries built using all available features score highly on helping users decide if they want to see full presentations. The purpose of summaries built using a subset of available features was to evaluate the effectiveness of individual features.

The results of eye-tracking experiments performed in this study indicate that 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 areas outside of the attention zone during summaries than during full presentations. This can be seen from Table 2 and Figures 1, 2, 3 and 4.

Table 3: Questionnaire Results - Likert scale.

Video	Q1	Q2	Q3	Q4	Q5
plen2_Classify	2.625	3.75	3.125	3.4375	4.625
plen2_audio_only	2.3125	3.5	2.4375	3.8125	4.8125
plen2_video_only	3	4.625	2.4375	4.0625	4.75
plen2_no_keywords	3.5625	4.8125	3.75	4.25	5.0625
prp1_Classify	2.875	4.3125	2.3125	3.8125	4.5
prp1_audio_only	2.25	3.875	2.4375	2.9375	4.9375
prp1_video_only	2.375	3.25	2	2.875	4.0625
prp1_no_keywords	2.5625	3.8125	2.8125	3.875	5.0625
prp5_Classify	4.25	4.875	4.5	4.4375	4.8125
prp5_audio_only	3.625	4.25	3.375	3.625	5.125
prp5_video_only	4.25	4.5	4.4375	4.125	5.3125
prp5_no_keywords	5.0625	5.0625	3.8125	4.5625	4.875
spRT6_Classify	2.875	4.125	2.625	3.875	5.4375
spRT6_audio_only	2.8125	4.5625	2.5	4	5.0625
spRT6_video_only	2	3.5	2.5	3.0625	5.125
spRT6_no_keywords	2.6875	3.9375	2.4375	3.3125	4.5

Table 4: 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.

6 CONCLUSIONS

This paper describes our investigation into the summarisation of academic presentations using linguistic and paralinguistic features. Comprehensive evaluations of summaries are reported including eye-tracking and the development of summaries using subsets of available features and a questionnaire evaluation of these to discover the effects of individual classification features on final summaries.

The results of this study indicate that classification of areas of engagement, emphasis and comprehension is useful for summarisation. Although the extent of its usefulness may depend on how engaging and comprehensible presentations were to begin with. Presentations rated as not engaging tend to see bigger improvements in engagement levels of summaries than presentations already rated as highly engaging.

Gaze plots show large improvements for summaries. Results show increased fixation counts with reduced fixation durations for summaries, confirming that users are more attentive for presentation summaries. This difference is more pronounced for videos not already classified as highly engaging videos, backing up other results showing that the summarisation process is more affective for videos which have not already been classified as most engaging. Our earlier studies (Curtis et al., 2017b) also support these new results reported in this paper.

Questionnaire results on summaries built using a subset of features show that audio-only classifications and visual-only classifications result in summaries which are rated less easy to understand and less informative than summaries built using full information and with no key words. Summaries built using no keywords also lack coherence. Overall, these results are very promising and demonstrate the effectiveness of this automatic summarisation strategy for academic presentations.

In future work we aim to develop a conference portal where user's can view presentation summaries developed on the fly using the features described in this paper. This portal will then allow for further evaluation of the effectiveness of these features over a greater pool of participants. Future work will evaluate the effectiveness of using more linguistic features for summarisation.

ACKNOWLEDGEMENTS

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University.

The authors would like to thank all participants who took part in these evaluations. We would further like to express our gratitude to all participants who took part in our previous experiments for classification of the high-level paralinguistic features discussed in this paper.

REFERENCES

Chechik, G., Ie, E., Rehn, M., Bengio, S., and Lyon, D. (2008). 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.

Curtis, K., Jones, G. J., and Campbell, N. (2015). Effects of good speaking techniques on audience engagement. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 35–42. ACM.

Curtis, K., Jones, G. J., and Campbell, N. (2016). Speaker impact on audience comprehension for academic presentations. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 129–136. ACM.

Curtis, K., Jones, G. J., and Campbell, N. (2017a). 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.

Curtis, K., Jones, G. J., and Campbell, N. (2017b). Utilising high-level features in summarisation of academic presentations. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 315–321. ACM.

He, L., Sanocki, E., Gupta, A., and Grudin, J. (1999). Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 489–498. ACM.

Huiskes, M. J., Thomee, B., and Lew, M. S. (2010). 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.

Joho, H., Jose, J. M., Valenti, R., and Sebe, N. (2009). Exploiting facial expressions for affective video summarisation. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 31. ACM.

Ju, S. X., Black, M. J., Minneman, S., and Kimber, D. (1998). Summarization of videotaped presentations: automatic analysis of motion and gesture. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):686–696.

Lew, M. S., Sebe, N., Djeraba, C., and Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):1–19.

Pavel, A., Reed, C., Hartmann, B., and Agrawala, M. (2014). Video digests: a browsable, skimmable format for informational lecture videos. In *UIST*, pages 573–582.

Rayner, K. and Sereno, S. C. (1994). Eye movements in reading: Psycholinguistic studies. *Handbook of psycholinguistics*, pages 57–81.