Development of an Annotated Multimodal Dataset for the Investigation of Classification and Summarisation of Presentations using High-Level Paralinguistic Features

Keith Curtis[⋆], **Nick Campbell**[†], **Gareth J. F. Jones**[⋆],
ADAPT Centre[⋆†]
School of Computing[⋆], School of Computer Science & Statistics[†]
Dublin City University[⋆], Trinity College Dublin[†]
Dublin, Ireland
Keith.Curtis@dcu.ie, nick@tcd.ie, Gareth.Jones@dcu.ie

Abstract

Expanding online archives of presentation recordings provide potentially valuable resources for learning and research. However, the huge volume of data that is becoming available means that users have difficulty locating material which will be of most value to them. Conventional summarisation methods making use of text-based features derived from transcripts of spoken material can provide mechanisms to rapidly locate topically interesting material by reducing the amount of material that must be auditioned. However, these text-based methods take no account of the multimodal high-level paralinguistic features which form part of an audio-visual presentation, and can provide valuable indicators of the most interesting material within a presentation. We describe the development of a multimodal video dataset, recorded at an international conference, designed to support the exploration of automatic extraction of paralinguistic features and summarisation based on these features. The dataset is comprised of parallel recordings of the presenter and the audience for 31 conference presentations. We describe the process of performing manual annotation of high-level paralinguistic features for speaker ratings, audience engagement, speaker emphasis, and audience comprehension of these recordings. Used in combination these annotations enable research into the automatic classification of high-level paralinguistic features and their use in video summarisation.

Keywords: Data Collection, Annotation, Classification

1. Introduction

Online archives of presentations provide valuable sources of material for study and research. However, such is the amount of this content available in many settings it can be difficult for users to efficiently access material which is most likely to be of interest to them. Search of this content based on the words spoken enables potentially relevant material to be identified based on the topic being presented. However, text-based analysis does not support location of the most important or emphasised material on a particular topic as indicated by the behaviour of the speaker or the audience. Use of high-level paralinguistic features which form part of the presentation offers the potential to identify the most significant topically relevant material within a presentation. Research into the automatic identification of such paralinguistic features and their utilisation in guiding users to relevant material by the use of applications such as summarisation requires the development of suitable corpora to support this work.

Such a dataset must include the relevant audio visual content with suitable manual annotations of the features to be extracted. We describe the construction of such a corpus designed to support this research. Our corpus consists of recordings of paper presentations at the Speech Prosody 7 conference, an academic conference held in Dublin, Ireland in May 2014. The recorded contents include audio-visual content of the presenter, but also parallel recordings of the audience to each presentation.

To investigate the automatic identification of paralinguistic features, these recordings were manually labelled with regions of emphasis by the speaker, ratings of the effectiveness of the speaker, audience engagement with the presentation and their comprehension of its content.

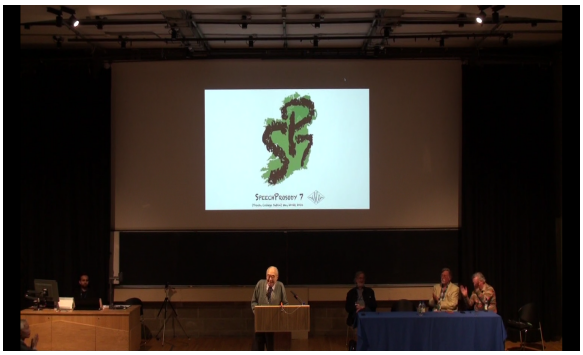

Figure 1: Full camera view of the stage

The remainder of this paper gives further details of the recorded data collection and the processes of its annotation with paralinguistic features.

2. Data Collection

The Speech Prosody 7 conference included a range of presentation types including keynotes, oral presentation of full papers and poster presentation. Our collection consisted of 31 full paper presentations. These were recorded in high quality with fully synchronised recordings of the audience to each presentation. Recordings have a full view of the stage, including the slides used for the presentation. In addition to this, a pdf version of the slides used in each presentation was archived. A total of three fixed cameras

Figure 2: Audience View

were used to record each presentation. Two cameras were fixed within the gallery facing the speaker at the approximate mid-point of the seating structure. One camera was set to record the overall wide-angle view of the whole stage, including the presenter, slides and the surrounding stage area, as seen in Figure 1. 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. The third camera was set up just behind and slightly to the side of the presenter in order to record the audience during each presentation, demonstrated in Figure 2.

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 academic presentations.

Recordings were made using three SONY HDR-XR500 cameras. Video was recorded in 1080p at 29.97 fps with an H.264 codec. Audio was recorded in Dolby Digital 48kHz, 16 bit stereo at 256 kbps. The recording standard used was AVCHD.

Presentation recordings were later synchronised by matching presenter recordings and audience recordings using the acoustic footprint. The recorded audio was then synchronised to the video recordings. Recordings were later processed by *Spoken Data* [1], to create full automatic speech transcripts and keywords for each presentation provided. Speech transcripts and keywords were provided by external processing by *Super Lectures* [2].

Videos were processed with an H.264 codec in mp4 format. These processed videos had a frame rate of 25 fps and a bit-rate of 768 kbps. Audio recordings were processed with MPEG-4 AAC audio codec with a sample rate of 44100 Hz and an audio bit-rate of 86 kbps.

The content of the presentations total 520 minutes of conference video, with an average presentation length of a little over 17 minutes. This gave a total of 1040 minutes of video for the presenter and audience recordings requiring human annotation for each paralinguistic concept studied.

3. Human Annotation

Human annotations of the recordings were made to create gold standard labels of the paralinguistic features. The annotated features were: emphasis, speaker ratings, audience engagement and audience comprehension. These features were expected to provide useful information in terms of identifying regions of the presentations which will be of most interest to users of the videos. Annotators were required to watch the content using a specially developed web-based annotation tool and to estimate levels of engagement and comprehension, or in the case of spoken emphasis, to estimate whether or not the content is emphasised. Presentation videos were uploaded to *YouTube*, from which video segments were embedded into the annotation tool.

Annotators for 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. A total of 40 annotators were equally balanced between native English speakers and non-native speakers. Some, but not all annotators, had prior experience of working with spoken content. Annotation records from each annotator were captured which later allowed us to analyse ratings made by individual annotators. Annotators for comprehension were recruited using a popular crowdsourcing website, further details are given within the description of the comprehension annotation procedure in Section 3.3.

3.1. Speaker Ratings & Audience Engagement

Our annotations made use of the scheme for annotation of conversational engagement provided in (Bednarik et al., 2012). This labels conversational engagement over 6 levels of engagement for window size of 15 seconds. Since the annotation requirements for our dataset differ from their work in that we are marking general audience engagement as they follow an ongoing presentation, we designed our annotation scheme in a similar manner, though using different levels of engagement and window sizes due to the differences in the annotation task.

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, in order to match speaker ratings and audience engagement levels for corresponding video segments.

We asked each annotator 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 engage-

[1] https://spokendata.com/

[2] https://superlectures.com/

Figure 3: Presenter Close-up view

ment levels. Annotators were also requested to provide an estimate of the attendance level at each talk, i.e. how full was the auditorium, estimated on a scale from 1 to 5.

Following this annotation task, participants were asked to watch 30 second video segments of the speaker, 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 were provided on an ordinal 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. Even numbered ordinal scales were chosen for this evaluation in order to force annotators into making a decision rather than selecting the middle option. 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.

Each video segment was annotated once, thus further steps were taken to eliminate bias in annotation as explained below. For annotation we followed the assumption, as observed from watching the dataset, that audience engagement levels do not vary much over a short period of time. As each segment was annotated just once for this concept, a number of steps were taken to prepare the data in order to reduce annotator bias and to smooth the annotations, since it was observed that individual annotators can have a tendency to annotate on the high side, or on the low side, when performing annotation tasks.

Outlier Removal The first step involved the removal of obvious outliers from the dataset. Outliers were defined as labels which did not match well with nearby segment annotations. An example would be where a sequence of video segments received an engagement rating of 4, followed by a segment receiving an engagement rating of 1. Outliers were removed to be re-annotated by different annotators from the pool.

Normalisation The next step involved the normalisation of labels to account for annotator bias. This was achieved by analysing ratings for each annotator and applying either a lowering or raising of annotator labels to bring each annotator's ratings in line with other annotations. This was necessary since 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 smooth annotations and reduce the effect of annotator bias. Video segments were aligned into time windows each 90 seconds in length, i.e. combining three consecutive segments 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.

3.2. Emphasis

The next task was to obtain human annotations for intentionally or unintentionally emphasised speech in our presentations. For annotation of these emphasised parts of audio-visual presentation, we first asked human annotators to watch two five minute clips from an audio-visual presentation and to mark areas of the video where they perceived 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, 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 enable investigation of potential correlations between emphasised speech and audience engagement.

Annotation of emphasis over the audio stream only was reported in (Kennedy and Ellis, 2003), where annotators listened to 22 minutes of speech audio and marked utterances which they considered to be emphasised. Annotators were provided with speech transcripts while listening to the audio stream, and marked emphasised points on the transcripts for all sentences they considered emphasised. Since the annotation required in our task differed from this earlier work, in that we are studying emphasis over the audio and visual streams, which had not been studied before, we showed the visual stream to the annotators as they listened to the speech. We considered there to be no need to give annotators the full transcripts as context should already be available to annotators from the visual stream.

A total of 10 annotators were recruited for our speech emphasis annotation task, and were paid 5 euro each after completion of tasks. First, 5 annotators labelled emphasis for two five-minute presentation segments, and engagement levels of the audience for another two five-minute presentation segments. The other five annotators labelled audience engagement for the first two videos and emphasis for the final two videos to ensure that each of the four video segments chosen for this task was labelled for emphasis by 5 annotators and the audience for engagement by five different annotators. This ensured that no annotator labelled audience engagement for the same video that they had already annotated for emphasis. This was to prevent bias during annotation of engagement levels which may occur if the presenter was already aware that they may have labelled for

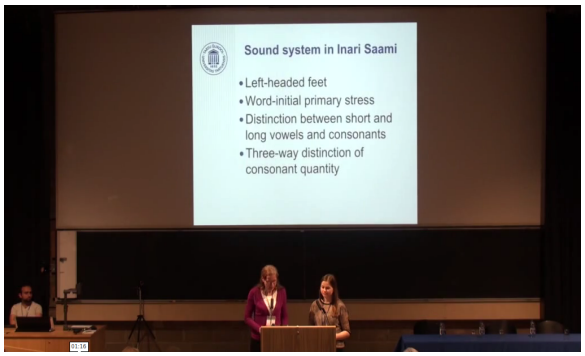

Figure 4: Two Presenters jointly present a talk

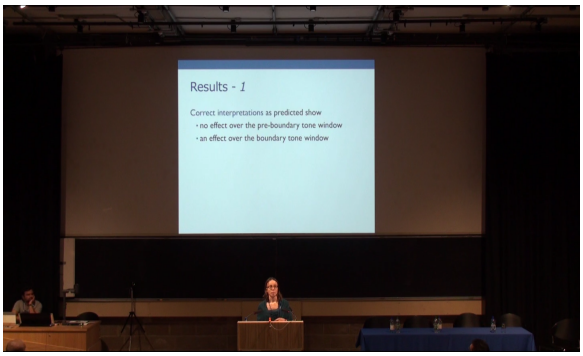

Figure 5: Annotators View

emphasis at a particular point in time.

Following this initial annotation for emphasised speech, it was clear that high levels of disagreement existed between annotators. We considered this to be due to the high level of subjectivity on just what constitutes emphasised speech, and meant that the training of a standard classifier on this data to classify emphasised speech was not practical. Upon further study of areas of agreed emphasis between annotators, a set of potential conditions to satisfy emphasis were constructed and an algorithm developed to identify all possible areas of potentially emphasised speech. All potential areas of emphasis identified by this algorithm were later judged by three human annotators. These sections of the speech data were marked as either emphasised or not emphasised. The mean intra-class correlation between these judgements was calculated as 0.5818, giving us a good level of inter annotator agreement between judges.

3.3. Comprehension

No study of audience comprehension has been reported previously, meaning no existing annotation schemes for tasks of this nature were available to us. Thus, we designed our own annotation scheme based on the information available to prospective annotators.

In order for our dataset to be able to support the study of the concept of audience comprehension within the presentations, we needed to create gold-standard labels for comprehension levels over the dataset. For this task each of the presentations in the dataset was divided into between 4 and 7 contiguous video segments. Each video segment was between 2 and 4 minutes in length. This gave a total of 172 video segments requiring manual annotation.

Annotators were asked to watch each video segment, in order, and to provide a short, written summary of the presentation segment. The purpose of these written summaries was to have them think about the content first before providing their comprehension estimate and also to provide for a method 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 the annotators to make a definite decision on comprehension level rather than choosing a middle, neutral option.

An alternative to using video in this way might be to have required audience members from the presentation to provide information of their perception of the comprehensibility its when attending it. In the absence of this information we consider our annotation method the best approach available.

As stated at the beginning of this section, the comprehension annotators were recruited from a popular crowdsourcing website. Annotators were paid an average rate of 7.50 euro per hour. Recruited human annotators all had English as their first language and all had at least some level of university 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. The quality of their work was checked before payment was made by studying their provided text summaries and comparing them with their estimated levels of comprehension. For example, if an annotator was unable to provide an accurate text summary of the presentation they had just watched, then it is unlikely they could have a high level of comprehension for that segment. Thus, any high level of comprehension reported by the annotator for that segment could not be taken as reliable.

To calculate 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 the 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* [3]. The mean ICC(1,1) score was found to be 0.6034, which considering the subjectivity of the task at hand we regard as a good level of agreement between judges.

4. Concluding Remarks

In this paper we have described the creation of a multimodal dataset of academic presentations from an international conference and its annotation for the investigation of high-level paralinguistic features to support access to regions most likely to be of interest to a user of this content. While we are unable to release our collection to the wider

[3] http://department.obg.cuhk.edu.hk/researchsupport/IntraClass_correlation.asp

research community, the methods described can be applied to other collections.

The annotation methods described in this paper have proven effective in our work, as shown in the following publications emanating from this work. We demonstrated effective classification of audience engagement in (Curtis et al., 2015). We showed annotation for the identification of emphasised regions of speech to be effective in (Curtis et al., 2017a). We demonstrated classification of audience comprehension during academic presentations to be effective in (Curtis et al., 2016). Finally, we demonstrated the usefulness of all these methods for the summarisation of academic presentations (Curtis et al., 2017b; Curtis et al., 2018), which highly depends on all of the features discussed in this paper.

5. 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. We would like to thank all of our annotators for participating in this study.

6. Bibliographical References

Bednarik, R., Eivazi, S., and Hradis, M. (2012). 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.

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*, pages 37–42. Linköping University Electronic Press, Linköpings universitet.

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.

Curtis, K., Jones, G. J. F., and Campbell, N. (2018). Summarising academic presentations using linguistic and paralinguistic features. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: HUCAPP*, pages 64–73. INSTICC, SciTePress.

Kennedy, L. S. and Ellis, D. P. (2003). Pitch-based emphasis detection for characterization of meeting recordings. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 243–248. IEEE.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.