Identification of Emphasised Regions in Audio-Visual Presentations

Keith Curtis
ADAPT Centre
School of Computing
Dublin City University
Ireland
Keith.Curtis
@AdaptCentre.ie

Gareth J.F. Jones
ADAPT Centre
School of Computing
Dublin City University
Ireland
Gareth.Jones
@dcu.ie

Nick Campbell
ADAPT Centre
School of Computer
Science & Statistics
Trinity College Dublin
Ireland
nick@tcd.ie

Abstract

Rapidly expanding archives of audio-visual recordings available online are making unprecedented amounts of information available in many applications. New and efficient techniques to access this information are needed to fully realise the potential of these archives. We investigate the identification of areas of intentional or unintentional emphasis during audio-visual presentations and lectures. We find that, unlike in audio-only recordings where emphasis can be located using pitch information alone, perceived emphasis can be very much associated with information from the visual stream such as gesticulation. We also investigate potential correlations between emphasised speech, and increased levels of audience engagement during audio-visual presentations and lectures.

1 Introduction

The rapidly expanding archives of audio-visual recordings available online are making unprecedented amounts of information available in many applications. However, realising the potential of this content requires the development of new and innovative tools to enable the efficient location of significant content of interest to the user. Manually browsing multimedia archives to identify audio-visual content of interest is extremely time consuming. While browsing of this sort of content is challenging for multimedia archives in general, it is an extremely challenging problem for archives of spoken content where most of the information exists in the audio stream.

Addressing requires the development of new tools to allow the user to search for potential content of interest without having to first listen to it. In our current work we are interested in identification of areas of speaker emphasis in audio-visual presentations. This has the potential to improve applications such as automatic summarisation or browsing of audio-visual content. Tools of this nature could also potentially be used for improved search and retrieval capabilities.

Previous work has explored identification of emphasised speech using the audio-only stream. In this work we expand on this earlier work in an audio-visual context to demonstrate that emphasis detection can be more successfully achieved using a multimodal analysis approach.

We also address the question of whether speech emphasis in the context of audio-visual presentations or lectures shows a correlation with what is typically referred to as ‘good’ public speaking techniques. This concept of ‘good’ public speaking techniques has been investigated in our previous work (Curtis et al., 2015). Given that emphasis is normally applied by the speaker to draw the attention of the audience to a specific part of speech for reasons of clarity or importance, we also investigate whether this applied emphasis affects change in the overall levels of engagement among the audience to such material.

2 Previous Work

Previous work (Chen and Withgott, 1992) has studied the use of emphasis for automatic summarisation of a spoken discourse. In this work emphasised speech from one speaker was detected and summarisation excerpts were extracted with no noticeable differences from human extracted summarisation excerpts. The data used was a 27 minutes videotaped interview between two primary speakers and the second was a set of phrases extracted from a telephone conversation. The emphasis model was trained on a

Hidden Markov Model (HMM) in which three separate models were created for 3 speech emphasis levels: emphatic speech, unemphatic speech and background speakers.

Another study (Arons, 1994) performed pitch based emphasis detection for automatic segmentation of speech recordings. In this work a pitch threshold of the top 1% of pitch values was chosen, speech segments with pitch values exceeding this threshold were classified as emphasised speech. From this, the pitch based segmentation technique was used to summarise the speech recordings into the most important speech segments. (He et al., 1999) attempted to summarise audio-visual presentations using pitch values in the top 1 percentile. In this work they found that audio-visual presentations were less susceptible to pitch based emphasis analysis than the audio stream only.

Following on from this work, (Kennedy and Ellis, 2003) studied emphasis detection for characterisation of meeting recordings. In this work 5 human annotators labelled 22 minutes of audio from the International Computer Science Institute (ICSI) meeting corpus. Annotators were given both an audio recording and a transcript from the meeting. Annotators listened to the audio recording while working their way down the transcript and marking each utterance as emphasised or not. They 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 agreed on emphasis accuracy rates of 92% were achieved. In addition, the utterances found to be the most emphasised were found by annotators to be a good summarisation of the meeting recording.

To the best of our knowledge, the detection of regions of speech emphasis has not previously been performed in an audio-visual context. (He et al., 1999) indicate that emphasis in audio-visual recordings is indicated by more than just notable increases in pitch as in the audio stream. In this study we investigate use of audio-visual features to detect emphasis in academic presentations. Also to the best of our knowledge, this concept has not been investigated for potential correlations with resulting audience engagement to the presentation or lecture material at hand.

3 Multimodal DataSet

For this study we used the International Speech Conference Multi-modal Corpus (SCMC) developed in our previous work (Curtis et al., 2015). This contains 31 academic presentations totalling 520 minutes of video, and includes high quality 1080p parallel video recordings for both the speaker and the audience to each presentation. Recordings have a frame rate of 29.97 fps, and were recorded at H264 codec. High quality parallel audio recordings are also included for each presentation in addition to close-up recordings of each presenter's slides. The majority of presentations are standard podium presentations by a single speaker in front of a seated audience. Two of the presentations consisted of podium presentations by two speakers in front of their seated audience. For this study, four presentations were selected from the corpus, two of which had male presenters and two with female presenters, each presentation was in English.

To limit the size of this preliminary investigation, a single 5-minute clip was selected from each presentation, totalling 20 minutes of presentation video used in this initial study. Segments were chosen to include presenters who were judged by human annotators in previous work on this dataset to be good presenters (Curtis et al., 2015), and to exclude regions of speech not of the presenter.

4 Multimodal Feature Extraction

For our investigation we extracted the following audio-visual features from the recorded presentations:

Pitch: AutoBi Pitch Extractor (Rosenberg, 2010). We use default min and max values of 50 and 400 respectively. Pitch values over the entire range were normalised for each speaker.

Intensity: AutoBi Intensity Extractor (Rosenberg, 2010). This generated an Intensity contour using default parameters of a minimum intensity of 75dB and a timestep of 100ms. Intensity values over the entire range were normalised for each speaker.

Head movement: OpenCV (Bradski and Kaehler, 2008) using Robust Facial Detection described in (Viola and Jones, 2004). For this task we used a Head and Shoulder cascade to detect the presenters head and return the pixel values for the location of the speakers head at that point in time. We then extracted

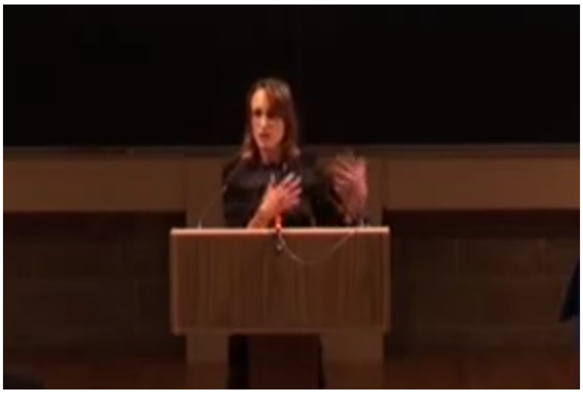

Figure 1: Presenter: mid-Emphasis

head movement by taking the Euclidean distance between pixel points in corresponding frames. These values were then normalised for each speaker.

Speaker Motion: was extracted using an optical flow implementation in OpenCV described in (Lucas et al., 1981). We calculated the total pixel motion changes from frame to frame to put more weight on directional changes in motion and take the mean and standard deviation in overall speaker motion. This accounted for change of direction in motion and represented variances of these values. These values were normalised per speaker.

5 Experimental Investigation

This section explores the experimental investigation we undertook during this research, including the human annotation of speaker emphasis during academic presentations.

5.1 Initial Investigation

The first part of this investigation involved asking a total of 10 human annotators to watch 2 of the 5 minute video clips taken from the four presentations, totalling 10 minutes of presentation video to be annotated for this concept. The annotators were asked to mark areas where they considered the presenter to be giving emphasis to speech, either intentionally or unintentionally. Due to the subjectivity of this task, annotators were instructed beforehand as to what exactly constituted emphasised speech, and were allowed to decide themselves just what they considered to be emphasised. There was however much disagreement between human annotators over areas of emphasis. We consider this to be due to the high level of subjectivity on just what it means to emphasise. This high-level of disagreement meant it was not practical to train a machine learning algorithm over this data for automatic classification of emphasised speech.

5.2 Further Investigation

Because of this disagreement, and in order to better understand the characteristics of regions consistently labelled as emphasised speech, we studied areas of agreed emphasis between the annotators. It was clear from this analysis, that consistent with earlier work, all agreed upon 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 features listed in the previous section to locate further candidate areas of emphasis.

The algorithm selects candidate regions by finding areas of high pitch in combination with areas of high motion or head movement. 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. 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 the top 20 percentile of gesticulation down to the top 40 percentile of values respectively. This resulted in the extraction of 83 candidate areas of emphasised speech from our dataset. These candidate regions were each judged for emphasis by three separate human judges, with the majority vote on each candidate emphasis region taken as the gold standard label for final agreement of emphasis.

6 Analysis and Results

Of the 83 candidate areas of emphasis extracted from presentation segments, 18 had pitch values in the top 1 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. The mean intra-class correlation 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, Thirteen 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.

7 Correlations Between Speaker Rankings and Emphasised Speech

We calculate prospective correlations between annotated speaker ratings and annotated emphasised speech. To achieve this we take values for 4 separate 5 minute video clips containing original emphasis annotations. We achieve 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 all of the 5 minute video clips combined gives a total of 32 time-windowed instances. We calculate correlations using the Pearsons Correlation Coefficient Calculator. Following this, we also calculate the correlation for speaker specific correlations between speaker ratings and emphasised speech. Table 1 outlines the results of these tests.

Table 1: Speaker Ratings - Emphasis : Linear Correlation

Video	*r* =
All_Combined	-0.3247
plenaryoral_2	-0.2988
plenaryoral_11	-0.0845
plenaryoral_12	-0.3362
prp_2	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 prp_2 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.

8 Correlations Between Audience Engagement Levels and Emphasised Speech

We also calculate prospective correlations between annotated audience engagement levels and annotated emphasised speech. Once again, to achieve this we take emphasis values for 4 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.

Calculating this over all of the 5 minute video clips combined gives a total of 32 time-windowed instances. Correlations are calculated using the Pearsons Correlation Coefficient Calculator. Following this, we also calculate the correlation for speaker specific correlations between audience engagement levels and emphasised speech. Table 2 shows the results, of which no clear correlation between these two concepts is visible.

Table 2: Audience Engagement - Emphasis : Linear Correlation

Video	*r* =
All_Combined	-0.1593
plenaryoral_2	-0.475
plenaryoral_11	0.2887
plenaryoral_12	0.8868
prp_2	0.1857

From Table 2 we can clearly see that correlation calculations per video appear to be very random, leading us to conclude that no correlations exist between audience engagement levels and emphasised speech. While the video plenaryoral_12 indicates a strong positive correlation, plenaryoral_2 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 carried out over a very small set of data.

9 Conclusions and Further Work

Previous work on emphasis detection in recordings of spoken content had looked at the concept in the context of the audio-stream only. Our small study shows 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. Further, no real correlations were discovered between areas of 'good' public speaking techniques or with audience engagement levels.

Previous work had discovered that emphasised speech can be used for effective summarisation of the audio-only stream (Chen and Withgott, 1992). Our future work will investigate the potential to summarise audio-visual lectures and presentations by using identified areas of intentional or unintentional speaker emphasis in addition to other paralinguistic features.

In this regard, initial experiments investigating the potential of identified areas of emphasised speech to be used for generating automatic presentation summaries have proven promising. In work combining identified areas of emphasised speech along with classifications for audience engagement and comprehension, early results have shown that generated summaries tend to be more engaging and information rich than full presentations, whilst participants tend to maintain focus for longer periods (Curtis et al., 2017).

10 Acknowledgments

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 human annotators for participating in this study.

References

Barry Arons. 1994. Pitch-based emphasis detection for segmenting speech recordings. In *International Conference on Spoken Langauge Processing*.

Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc.

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

Keith Curtis, Gareth JF Jones, and Nick Campbell. 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.

Keith Curtis, Gareth JF Jones, and Nick Campbell. 2017. 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.

Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. 1999. Auto-summarization of audio-video presentations. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, pages 489–498. ACM.

Lyndon S Kennedy and Daniel PW Ellis. 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.

Bruce D Lucas, Takeo Kanade, et al. 1981. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, volume 81, pages 674–679.

Andrew Rosenberg. 2010. Autobi-a tool for automatic tobi annotation. In *INTERSPEECH*, pages 146–149.

Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.