Thin slicing to predict viewer impressions of TED Talks

Ailbhe Cullen[1], *Naomi Harte*[1]

[1]Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
cullena3@tcd.ie, nharte@tcd.ie

Abstract

Many paralinguistic challenges have looked at predicting affect, speaker state, or other attributes from short segments of speech of less than a minute. There are situations however, where we want to predict how a user might label a talk or lecture of significantly longer duration. For example, would a viewer find a given talk funny? The question then is how to map long talks to single word labels? In this paper, we rely on the concept of thin slicing, which states that humans make similar judgements on short segments of speech as they do on longer segments. We wish to find short segments that are representative of the talk, which can be used to predict the user label. We explore this concept in order to predict user ratings of TED talks as inspiring, persuasive, and funny. In particular, we pose two questions. The first is how thin can we make our slices? Results show that longer slices, of up to a minute in duration are more useful for the prediction of viewer ratings. We also ask where the best position to slice the video is? We compare the performance of classification based on slices extracted from fixed points to that of slices extracted from salient regions, and find that prediction accuracy can be improved by choosing slices according to the speaker's vocal behaviour or the audience's reactions.

Index Terms: audio visual affect recognition, social signal processing, human-computer interaction, computational paralinguistics

1. Introduction

To date, a major focus in social signal processing has been on analysis and prediction of affect based on short segments of recordings. Pfister and Robinson [1], Weninger et al. [2], and Biel et al. [3] perform annotation of paralinguistic traits over approximately 60 sec clips. Strangert and Gustafson [4] use samples between 30 and 36 sec long, while Lubis et al. define a “Tri-Turn”, three consecutive sentences, as their unit of analysis [5]. D’Errico et al. [6] worked at a sentence level. This approach is useful for many tasks, but there are some situations where a human will engage with an entire lecture or on-line talk and judge it as funny, inspiring or boring, for example. In this study, we are interested in the mapping of long talks to single-word labels.

Psychological studies have shown that high-level speaker traits (such as personality or speaker appeal) can be effectively judged by human annotators from short slices of recordings [7, 8]. While the exact relationship between slice length and position, and rater accuracy is dependent on the specific phenomenon being rated, a general trend has been reported for accuracy to increase as slice length increases from 5 sec up to 60 sec [7]. Beyond this point, there is no further benefit to increasing slice length. This is the concept of thin-slicing. The optimal slice length remains an open question. In [9], Mariooryad et al. examined inter- and intra-speaker variability in short segments of video lectures, and concluded that a duration of 20 seconds was sufficient for judging speaking style. However, it appears this consistency was only demonstrated for a small portion of their database.

The TED talks vary in length, from 3 to 30 minutes, but are assigned single word labels by raters, representing the overall impression of the talk. Thus, there is a large mismatch between the resolution of audio and video information compared to the resolution of the target labels. There is also a high level of variation in delivery style within talks. This raises the question of how best to process such realistic talks in an automatic system? Is it possible to somehow find the ”best” segments of a talk that represent the attributes of that talk, rather than having to process the entire talk? To explore this issue, our paper investigates the automatic classification of TED talks based on slices of 15, 30, and 60 seconds, extracted from different positions in the talk to predict whether a given talk has above-average ratings for three key-words: funny; inspiring; and persuasive. We compare the performance of different slicing regimes on a significant task size of 306 talks. We examine the impact of both the duration of the slice, and the position from which the slice is taken on the overall performance of the classifier.

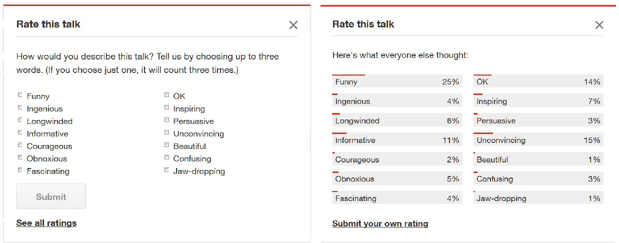
The remainder of the paper is arranged a follows: Section 2 introduces the TED Talks database, and discusses the user-generated labels which will be used in this study. In Section 3 we introduce the concept for thin sliing, discuss how slices are chosen, and outline the features extracted from each slice. The performance of this slicing technique is discussed in Section 5. A final conclusion is given in Section 6.

2. TED Talks as a crowd-sourced database

TED[1] is a non-profit organisation which hosts conferences and speakers from a wide range of disciplines. The TED organisation maintains an online library of talks from these conferences, which is available under creative commons for non-commercial use. A significant on-line community has developed around the TED talks. Users can create accounts on the TED website in order to comment on talks, save their favourites, create playlists, and rate talks using fourteen key-words. This information has been used for a number of semantic and paralinguistic tasks. Pappas et al. combine user references, talk meta-data such as descriptions and subject tags, and a semantic processing of comments to improve video recommendations [10,11]. More closely related to this study, Tsai [12] performs an analysis of prosody in TED speakers and university lecturers, in order to establish what characteristics make a good talk or lecturer. Finally, Salim et al. [13] use simple global features, such as the amount of times a speaker elicits applause and the portion of the video for which the speaker is in view, to predict the talk key-word ratings.

[1] http://www.ted.com/

10.21437/AVSP.2017-12



Rate this talk ✕

How would you describe this talk? Tell us by choosing up to three words. (If you choose just one, it will count three times.)

- □ Funny
- □ Ingenious
- □ Longwinded
- □ Informative
- □ Courageous
- □ Obnoxious
- □ Fascinating
- □ OK
- □ Inspiring
- □ Persuasive
- □ Unconvincing
- □ Beautiful
- □ Confusing
- □ Jaw-dropping

(a) *Rating Form*

Rate this talk ✕

Here's what everyone else thought:

Funny	25%	OK	14%
Ingenious	4%	Inspiring	7%
Longwinded	6%	Persuasive	3%
Informative	11%	Unconvincing	15%
Courageous	2%	Beautiful	1%
Obnoxious	5%	Confusing	3%
Fascinating	4%	Jaw-dropping	1%

(b) *Sample Results*

Figure 1: *TED rating form and results for talk id. "AJJacobs_2011P".*

Table 1: *Distributin of binary (High/Low) and tertiary (High/Mid/Low) labels for the 306 selected TED talks. H = High, M = Mid, L = Low.*

Label Type		Funny	Inspiring	Persuasive
Binary	H	0.21	0.38	0.43
	L	0.79	0.62	0.57
Tertiary	H	0.28	0.39	0.42
	M	0.18	0.21	0.21
	L	0.44	0.40	0.37

2.1. TED labels

In this study we use TED's community generated ratings as crowd sourced labels. On the TED website, viewers are given the option to rate talks using up to three of fourteen key-words: beautiful; confusing; courageous; funny; informative; ingenious; inspiring; jaw-dropping; long-winded; obnoxious; ok; persuasive; and unconvincing. An example of the rating form is given in Figure 1. Each viewer is given three votes. They may assign all three votes to a single word, or spread them across two or three words. In this chapter we focus on three labels: funny; inspiring; and persuasive. All three are chosen for their relevance to indexing and content retrieval.

We downloaded the raw vote counts for all fourteen labels through the TED API in January 2015. These were normalised by the overall number of votes for each recording. We then calculate the mean rating for each label by averaging across all talks which have a non-zero rating for this label. In the following experiments we perform a binary classification, to predict whether a talk is rated above or below average for a given label. This could potentially facilitate automated indexing of new uploads, or improved filtering of search results according to whether or not they are funny, inspiring, or persuasive. However, as can be seen in Table 1, the binary labels are not evenly distributed, particularly the funny label. This is likely to limit our ability to accurately predict the minority class. Thus, we also perform a three-level labelling, using the 40th and 60th quartiles as threshold points for assigning labels to the high, medium, and low classes.

2.2. Data Selection

When the TED library was crawled in January 2015 it contained over 2000 talks. This number is continually rising as more TED events are held, and more talks are uploaded to the website. With talks lasting up to 30 mins in length, it becomes costly to perform audio and video processing on all talks. Therefore we made a number of decisions in order to reduce the data set and thus the computational load of the following experiments.

While there are subtitle tracks available for all TED talks, these are not time aligned. Therefore, in order to enable future word or sentence level processing, we used only talks which were contained in the TED-LIUM database [14]. This reduces the database to 1495 videos, with a total duration of 207 hours. This is still a large amount of video to process. Thus we further reduced the dataset by choosing only talks between 6 and 12 mins in length, and discarding any performance-style talks (consisting of singing, dancing, karate, and magic tricks). This resulted in a set of 306 talks, which will be analysed in this paper.

3. Thin Slicing

While it is known that humans can accurately predict speaker traits from short slices of recordings [7, 8], the optimum slice length and position remains an open question [7, 9]. Olivola et al. find that humans make lasting appearance-based attributions in as little as 100 ms, while Carney et al. find that 60 s may needed to make reliable judgements of dyadic interactions [7]. In this paper we wish to explore whether summary labels, based on viewings of 6 - 12 mins of data, can be predicted from thin slices of TED talks. For each experiment, we train and test a Support Vector Machine (SVM) using only a single slice from each video. Acoustic and visual features for each video are extracted only from this single slice. In order to assess the effect of slice length we repeat each classification experiment using slices of 15, 30, and 60 seconds. The ultimate aim is to significantly reduce the computational load of processing large numbers of long videos.

Aside from the slice length, we also wish to determine the best position from which to take our slices. Thus we compare the performance of systems trained on a single slice extracted from either the start, middle, or end of a talk, to systems trained on a slice taken from a particularly salient region of the talk. We consider two approaches to identifying salient regions in the talks, one based on the speaker's prosodic behaviour and one exploiting audience feedback.

Previous studies have found strong links between pitch and speaker ability [1,2,15,16]. Therefore, for each talk, the pitch is extracted over the whole talk. We then measure the pitch range and variance over a sliding window of 15 - 60 seconds (according to the desired slice length). The slices with the highest pitch range and variance are chosen for further feature extraction.

The audience is often considered a noise source, something to be filtered out or compensated for. However, in the context of viewer perception of a speaker, the audience becomes an important source of information. Salim et al. use the number of laughter or applause bursts as a feature to predict TED talk ratings [13]. Similarly, Strapparava et al. use audience reactions to predict persuasion in political speeches [17]. The subtitles provided with each TED talk contain annotations of audience laughter and applause. We use the subtitles to locate the longest instance of laughter and applause in each video, and then perform feature extraction on the slice preceding this. We will compare three choices for slice position: the slice preceding the longest laughter burst; the slice preceding the longest applause burst; and the slice preceding the longest feedback burst (this may be laughter or applause).

4. Classifier System

Having chosen a slice according to the analysis outlined above, we then extract a number of acoustic and visual features from

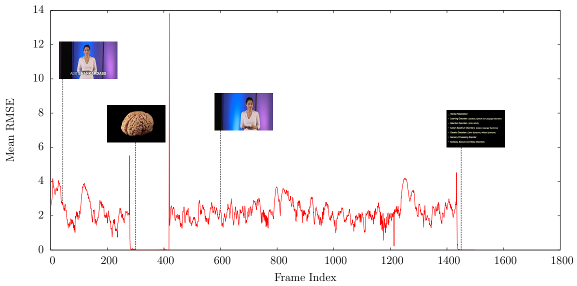


Figure 2: *RMSE calculated over the first 1500 frames (60 secs) of AdditiShankardass_2009I.mp4. Frames are displayed for indexes 43, 300, 600, and 1450, showing slides at position 300 and 1450.*

Algorithm 1 Hand Tracking

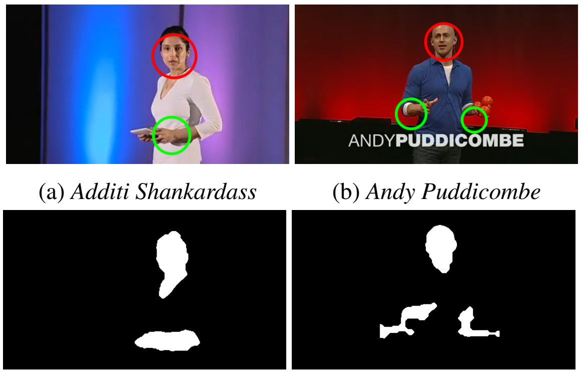
```
Get slice start and end indexes
for $i = f_{start}$ to $f_{end}$ do
    Read $i^{th}$ frame
    if frame $\neq$ slide then
        Viola-Jones Face Detection
        if face detected then
            Record face centroid and size
            Estimate skin tone from face
            Find skin tone in rest of frame
            Clean output using morphological operations
            Look for connected regions in mask
            Discard face region
            if num(handCandidates)$\geq$1 then
                if Hands visible in previous frame then
                    Choose candidate(s) closest to previous estimate(s)
                else
                    Choose candidate(s) with appropriate size and position relative to face
                end if
                Record hand centroids and sizes
            end if
        end if
    end if
end for
```

the slice. These are concatenated to form a single feature vector of length 1690. This is an order of magnitude larger than the size of our database (306 videos). To reduce over-fitting, we apply a principal component analysis (PCA) to the feature vectors. The top 50 highest weighted PCA components are used to train a linear SVM, to predict each of the three labels: funny; inspiring; and persuasive.

4.1. Visual Features

Hand and arm gestures are known to be important for both audience engagement [18] and emotion perception [19]. Tracking centroids of hands and arms has proven effective for sign language recognition [20]. There is a suggestion that this may be useful also for the detection of social signals [19, 21]. Thus in this section we present an algorithm to segment and track face and hand movement in the TED videos.

The hand tracking algorithm is outlined in Algorithm 1.



(a) *Additi Shankardass* (b) *Andy Puddicombe*

(c) *Additi Shankardass* (d) *Andy Puddicombe*

Figure 3: *Sample output of hand tracker, showing frames from (a) AditiShankardass_2009I.mp4 and (b) AndyPuddicombe_2012S.mp4 showing detected faces circled in red and hands circled in green, as well as the underlying skin tone detection (c,d).*

The TED videos often switch between showing the speaker and the speaker's slides. Thus, the first task is to detect when a slide is being displayed. We do this by calculating the root mean squared difference (RMSE) between subsequent frames. Consider two frames in a sequence, F_{n-1} and F_n, each of size $I \times J$. Let $F_n(i, j)$ denote the pixel in position $[i, j]$ in the n^{th} frame. We define the root mean squared error as follows:

$$RMSE_n = \sqrt{\frac{1}{IJ} \sum_{i,j} \left(F_n(i,j) - F_{n-1}(i,j)\right)^2} \quad (1)$$

Slides are stationary across multiple frames, thus after an initial impulse the RMSE drops to zero for the duration of the slide. This is illustrated in Figure 2, and allows us to easily detect when a slide is being displayed.

For each frame that is not a slide we use the Viola-Jones face detector [22], implemented in the Matlab signal processing tool-box, to locate the speakers face. This is a relatively robust algorithm, however it will occasionally fail if the speaker is too far from the camera, or has their back turned to the camera. If a face is not detected we skip to the next frame. The face detector may also return multiple faces when face-like-objects appear in the background slides. In this case we choose the face candidate closest to the face detected in the previous frame. Where there is no face detected in the previous frame, we choose the candidate ranked highest by the Viola-Jones algorithm.

The detected face region is used to estimate the skin tone, and also to restrict the hand search area. First the red, green, and blue channels are each normalised. Then the skin tone is estimated by averaging the normalised pixel values in the detected face region in each color channel. The search area is defined as an area four times the width of the face, and 5 times the height, centred on the face. These limits are chosen heuristically, as hands are rarely observed outside this area in the videos under analysis. Within the search region, we compute the error between each normalised pixel value and the skin-tone estimate. If the error is below a threshold of 0.1 (chosen by inspection) then the pixel is assumed to contain skin. Finally, some morphological operations are used to clean the skin-tone estimation.

Examples of the face and skin-tone detection are given in Figure 3. Figures 3a and 3b show the original frame with the hand-tracking results overlaid. Heads are circled in red, and hands in green. Circles are centred on the detected centroid, and the radius is scaled according to the size of the detected hand/face area. In general the face detection is quite accurate. The skin color mat shown in Figure 3d shows three clear regions, one where the face is located, and two in the region of the hands. some of the area behind the speaker's hands have also been detected due to the similarity in colour between the red backdrop and the estimated skin-tone. In Figure 3c, the speaker's hands are held together, causing them to appear as a single object in the mat.

From the binary skin-tone image, the size (in pixels) and centroid(s) of the detected hand(s) are recorded for each frame. The centroid positions of the hands are matched to the estimates from the previous frame to decide which is left and right. In cases where more than two hand regions are detected the two closest to the previous frame's hands are chosen. Temporal smoothing is applied to hand and face centroid and size estimates using a median filter over a five frame window. The first and second order differences in centroids (i.e. speed and acceleration of motion) and size (to capture zoom effects) were then calculated. This gives us a set of basic frame level features. Mean and peak values were calculated for each basic feature, over all frames within the slice to generate a final slice-level feature vector.

4.2. Acoustic Features

We use the feature set proposed for the Interspeech 2010 paralinguistic trait challenge [23]. This consists of 1582 features and contains statistical, spectral, and prosodic features. These features have since been used as a baseline for a range of paralinguistic recognition tasks [2, 24, 25]. This feature set will be referred to as the IS2010 feature set hereafter. The advantage of using these features is two fold. Firstly, it is well documented that prosodic and spectral features capture affective content of speech [2, 26, 27]. Secondly, the use of standard, easily reproducible features, eases comparison between studies.

Low level IS2010 features are extracted over 25ms frames. Statistical functionals are then extracted from these base features over the length of the chosen slice (15 - 60 sec), resulting in a single, slice-level feature vector. Included in the statistical functionals are the range and variance of the pitch track. These are used to select two of the candidate slice positions: the slice with the highest f0 range; and the slice with the highest f0 variance.

5. Results

For evaluation, our 306 TED videos are spilt into three folds, each containing 102 videos. 3-fold cross validation is performed, in which two folds are used for training, and one for testing. The results reported in the following section are the average performance on the three folds. Binary classification performance is reported in Figure 4, while the accuracy of the three class prediction is give in Figure 5.

5.1. Slice Length

The effect of slice lengths varies across labels. The most consistent trends in slice length can be seen in the classification of whether or not a video is persuasive. Both binary (Figure 4) and three-way (Figure 5) classification of persuasive benefit from longer slice lengths, either 30 or 60 seconds. When predicting whether or not a talk is funny, the best performance on the three class problem is again most often achieved using a longer slice. The performance on inspiring is more mixed. In Figure 4 shorter slices give better accuracy on inspiring, for the majority of slice positions. However, when it comes to three-way classification, inspiring is best predicted using longer slices.

It is more difficult to draw conclusions from the binary classification results for funny. In this case, there appears to be very little effect to either changing the slice position or length. This is most likely due to the severely imbalanced class distribution for this label. In Table 1 we reported that only 21% of the talks in our dataset have above average ratings for funny. This has introduced a significant bias in the trained classifiers, which has a greater effect on classification performance than any variation of slice length or position.

5.2. Slice Position

The effect of slice position is similarly dependent on the individual speaker trait being predicted. For the prediction of funny, we had expected that the slice preceding the longest burst of laughter would give the highest accuracy. However, the results in Figure 5 show that the best slice to use for the three-way classification of funny is the final 30 seconds of the video. This may be because this is where the speaker delivers the "punch-line" of their story. It should be noted that slices chosen to precede the longest burst of applause or audience feedback provide similar, if slightly lower, performance. Thus, our initial assumption that the audience response is a predictor of funny talks may contain some truth.

For the classification of inspiring talks, there is a clear benefit to choosing slices according to audience applause, which holds for both binary and three class tasks. Higher performances are also achieved when using the end slice of the talk. The end slice is also the most effective slice for the classification of funny talks. It is possible that this because the final moments of the talk are where the speaker drives home their message. However, it is also possible that the end slice is most fresh in the viewer's mind when rating the talk, and thus influences their rating more than other slices.

Whether or not a speaker is persuasive is best classified using slices chosen according to pitch statistics. For binary classification of persuasiveness, the highest performing slice is the slice with the greatest f0 variation, while for three-way classification the slice with the highest f0 range is more effective.

6. Conclusions

Vocal behaviour, facial expression, gestures, and body language all play a part both in communicating emotion and speaker traits [19, 20, 28], and in maintaining audience interest and engagement [18, 29]. While it is certain that human viewers make lasting judgements based on short observations [7, 8], there has been little interest in the machine learning community in using such short slices to predict overall impressions of much longer videos. In this study we present a multi-layered analysis of videos, first using a combination of meta-data and pitch analysis to locate salient regions within the videos, and then performing audio-visual classification based only on the most informative exert, i.e. a single slice of duration between 15 and 60 seconds. We do this using a large, fully realistic database, which has been annotated by a global on-line community. This progression mirrors a growing awareness in the affective computing community that in order to develop systems which can solve real-world problems, we must be realistic about the quality of the data that we work with [2, 26].

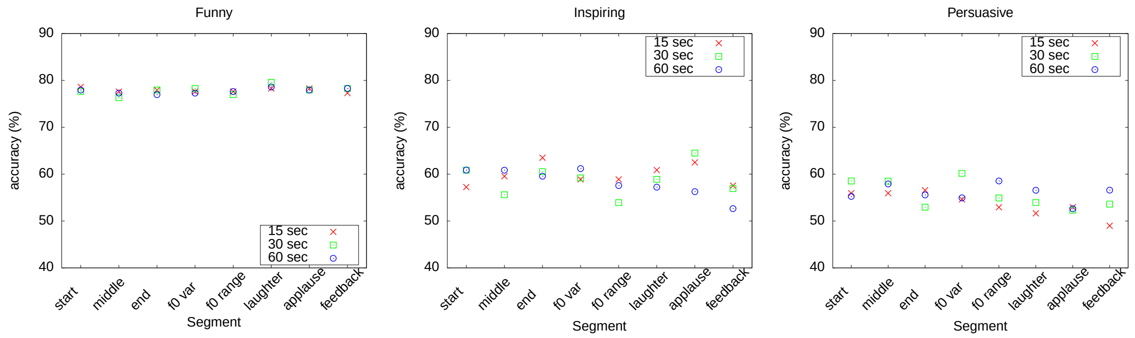


Figure 4: *Accuracy (% correct) of binary predictions of three TED labels.*

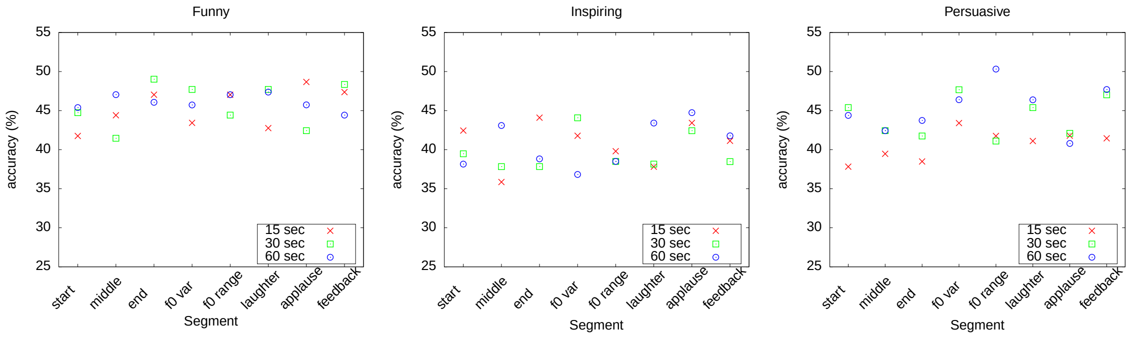


Figure 5: *Performance of three class predictions of TED labels.*

In the context of speech and emotion recognition, the term "multimodal" typically implies a system using a mixture of audio, video, and occasionally linguistic features [2,30–32]. However, in this section we presented a system which combines acoustic and video information in a new fashion, by first using the acoustic features or talk meta-data (subtitles) to locate important slices, and then performing visual analysis on only these slices. This allows us to make more efficient use of our resources, performing the simpler task of pitch detection on the entirely of talks, but only performing the more intensive visual processing on a small subset of the overall talk. Thus we exploit both modalities while reducing the overall computational load.

The best overall system for the prediction of funny is the one which uses a slice chosen according to audience laughter, and combines both the IS2010 and hand and face features at the classifier stage. Classification of inspiring videos is best using end slices and the same combination of IS2010 and hand and face features. The MFCC and hand and face features give the best performance on persuasive videos, using the slice chosen according to audience applause. Future work will focus on improving the visual features in order to further improve classification performance. There are a number of more elegant solutions for motion and gesture tracking available [19, 20]. Given the promising results obtained using relatively simple features and fusion techniques, we expect these improvements to give significantly increased performance.

In this paper, we compare the effect of varying slice length on classification performance, and find that in general, linger slices provide higher accuracies. In both Figure 4 and Figure 5 the absolute highest performance on each label is given by a 30 or 60 second slice. Thus, we conclude that longer slices contain more discriminative information. Across binary and three class tasks, we find that the best slice position for funny is he end slice, for inspiring is the slice preceding the largest applause burst, and for persuasive is the slice with the highest pitch variation or range. Beyond this, the end slice gives competitive performance on both the funny and inspiring prediction task. The end of a talk is typically where a speaker summarises and emphasises their message, thus we suggest that it has a strong and lasting impact on the overall perception of the talk.

7. Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

8. References

[1] T. Pfister and P. Robinson, "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," *Affective Computing, IEEE Trans.*, vol. 2, no. 2, pp. 66–78, 2011.

[2] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller, "The voice of leadership: Models and performances of automatic analysis in online speeches," *Affective Computing, IEEE Trans.*, vol. 3, no. 4, pp. 496–508, 2012.

[3] J. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *Multimedia, IEEE Trans.*, vol. 15, no. 1, pp. 41–55, 2013.

[4] E. Strangert and J. Gustafson, "What makes a good speaker? subject ratings, acoustic measurements, and perceptual evaluations," in *Interspeech*. ISCA, 2008, pp. 1688 – 1691.

[5] N. Lubis, S. Sakti, G. Neubig, K. Yoshino, T. Toda, and S. Nakamura, "A study of social-affective communication: Automatic prediction of emotion triggers and responses in television talk shows," in *Automatic Speech Recognitino and Understanding (ASRU)*, 2015, pp. 777 – 783.

[6] F. D'Errico, R. Signorello, D. Demolin, and I. Poggi, "The perception of charisma from voice: A cross-cultural study," in *Affective Computing and Intelligent Interaction (ACII), Humaine Association Conference on*, 2013, pp. 552–557.

[7] D. R. Carney, C. R. Colvin, and J. A. Hall, "A thin slice perspective on the accuracy of first impressions," *J. Research in Personality*, vol. 41, no. 5, pp. 1054–1072, 2007.

[8] C. Y. Olivola and A. Todorov, "Elected in 100 milliseconds: Appearance-based trait inferences and voting," *J. Nonverbal Behavior*, vol. 34, no. 2, pp. 83–110, 2010.

[9] S. Mariooryad, A. Kannan, D. Hakkani-Tur, and E. Shriberg, "Automatic characterization of speaking styles in educational videos," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4848–4852.

[10] N. Pappas and A. Popescu-Belis, "Combining content with user preferences for ted lecture recommendation," in *Content-Based Multimedia Indexing (CBMI), 11th Int'l Workshop*, 2013, pp. 47–52.

[11] ——, "Sentiment analysis of user comments for one-class collaborative filtering over ted talks," in *Proc. 36th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*. 2484116: ACM, pp. 773–776.

[12] T. Tsai, "Are you ted material? comparing prosody in professors and ted speakers," in *Interspeech*. ISCA, 2015, pp. 2534 – 2538.

[13] F. A. Salim, K. Levacher, O. Conlan, and N. Campbell, "Examining multimodal characteristics of video to understand user engagement," in *Conf. on User Modelling, Adaptation, and Personalisation (UMAP)*, 2015.

[14] A. Rousseau, P. Deléglise, and Y. Estéve, "Enhancing the tedlium corpus with selected data for language modeling and more ted talks," in *Language Resources and Evaluation (LREC), Int'l Conf. on*, 2014.

[15] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Sp. Comm.*, vol. 51, no. 7, pp. 640–655, 2009.

[16] P. Touati, "Prosodic aspects of political rhetoric," in *ESCA Workshop on Prosody*, 1993, pp. 168 – 171.

[17] C. Strapparava, M. Guerini, and O. Stock, "Predicting persuasiveness in political discourses," in *Language Resources and Evaluation (LREC), Int'l Conf. on*, 2010, pp. 1342 – 1345.

[18] J. R. Zhang, J. Sherwin, J. Dmochowski, P. Sajda, and J. R. Kender, "Correlating speaker gestures in political debates with audience engagement measured via eeg," in *Proc. 22nd ACM Int'l Conf. on Multimedia*. 2654909: ACM, pp. 387–396.

[19] M. Karg, A. A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, "Body movements for affective expression: A survey of automatic recognition and generation," *Affective Computing, IEEE Trans.*, vol. 4, no. 4, pp. 341–359, 2013.

[20] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer vision and image understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[21] T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer, "Multimodal public speaking performance assessment," in *Proc. 2015 ACM Int'l Conf. on Multimodal Interaction*. ACM, pp. 43–50.

[22] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[23] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Interspeech*. ISCA, 2010, pp. 2794 – 2797.

[24] W. Y. Wang and J. Hirschberg, "Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning," in *SIGDIAL*. Association for Computational Linguistics, 2011, pp. 152–161.

[25] S. Steidl, T. Polzehl, H. T. Bunnell, Y. Dou, P. K. Muthukumar, d. Perry, K. Prahallad, C. Vaughn, A. W. Black, and F. Metze, "Emotion identification for evaluation of synthesized emotional speech," in *Speech Prosody*, 2012, pp. 661–664.

[26] B. Schuller and A. Batliner, *Computational Paralinguistics*. Wiley, 2014.

[27] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech*, 2012.

[28] L. C. De Silva and N. Pei Chi, "Bimodal emotion recognition," in *Automatic Face and Gesture Recognition, Proc. 4th IEEE Int'l Conf.*, 2000, pp. 332–335.

[29] M. A. Nicolaou, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE Int'l Conf.*, pp. 1522–1526.

[30] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *J. Multimodal User Interfaces*, vol. 3, no. 1, pp. 33–48, 2010.

[31] Z. Zhihong, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Trans.*, vol. 31, no. 1, pp. 39–58, 2009.

[32] B. Schuller, R. Mller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.