Exploration of Judicial Facial Expression in Videos and Transcripts of Legal Proceedings

A thesis submitted for the degree of

Bachelor of Commerce (Honours)

by

Huize Zhang

27478343

Department of Econometrics and Business Statistics

Monash University

Australia

September 2019

Contents

Chapter 1

Introduction

Decisions by courtroom Justices have been discussed broadly in the legal literature. Gender, political views and religious background of both the Justices and counsel in the case potentially influence the decisions. This paper will explore the facial behaviour of the Justices during hearings with the objective of being to assess whether it can help to predict outcomes. Audio Visual(AV) recordings and case transcripts will be computationally processed and analysed to examine the decisions of each Justice.

1.1 Motivation

Tutton, Mack, and Roach Anleu (2018) attempted to utilize the AV technology, which is made available online by the High Court of Australia (Australia, 2019). They visually inspect of the videos to highlight when Justices depart from the expected norms of judicial conduct. To better understand the emotion status and therefore the departure of the emotional behaviour, more advanced technologies could be applied. An example is to use OpenFace (2018) technology, which provides information on emotions exhibited by the Justices. This technique has been applied by Kovalchik and Reid (2018) on professional tennis players during Grand Slam matches. That study demonstrated the potential to predict the outcome of High Court appeals based on Justices' demeanour utilising contemporary tools and emotion tagging techniques.

1.2 Literature review

The literature summary is divided into two parts: (1) current work in legal studies to understand the behaviour of the Justices and (2) existing facial recognition and emotion tagging technology.

1.2.1 Legal study from a behaviour perspective

People have attempted to predict the decisions of the Justices for centuries. Chief Justices of Australia and Zealand (2017) present the following code of conduct:

> It is important for judges to maintain a standard of behaviour in court that is consistent with the status of judicial office and does not diminish the confidence of litigants in particular, and the public in general, in the ability, the integrity, the impartiality and the independence of the judge.

This impartiality should be clear in judicial demeanour (Tutton, Mack, and Roach Anleu, 2018; Goffman, 1956). Paul Ekman (Ekman et al., 1991) suggests that from a behavioural perspective, some facial and vocal inflections are often unbeknown to the speakers themselves . Many scholars have exploited this in studying the court outcomes through the language and words used by the Justices in the court (Shullman, 2004) and vocal and facial characteristics of the Justices (Chen et al., 2018).

There are also existing works to understand the emotion of the Justices from a linguistic perspective and suggest some factors that could be useful to indicate how the Justices' vote and thus the court outcome. These factors include the use of pleasant and unpleasant language by Black et al. (2011), the frequency and content of Justices' questions by Shullman (2004) and Johnson et al. (2009). Epstein, Landes, and Posner (2010) use a regression analysis with the number of questions asked by the Justices used to infer the winning party in a case.

Other scholars (Chen, Halberstam, and Alan, 2016; Chen, Halberstam, Yu, et al., 2017; Schubert et al., 1992) have studied the emotion of the Justices from vocal characteristics and suggest that these vocal characteristics, especially perceived masculinity is strongly correlated with the court outcomes. Dietrich, Enos, and Sen (2019) uses a multilevel logistic model with random effects to suggest that subconscious vocal inflections contain information that is not available from text.

Chen (2018) employed both vocal and facial characteristics to predict the court votes using Supreme Court data from 1946-2014. The audio clips are first preprocessed to get the Mel-frequency Cepstral Coefficients (MFCC) and then applied to a random forest model. The image features are extracted using a Histogram of Oriented Gradients (HOG) method. More specific facial recognition software is readily available to extract human facial features and these facial recognition technologies have not yet been applied to the legal proceedings.

Most of the literature is conducted using the U.S. Supreme Court Database and less studies have been conducted using Australian High Court data. Tutton, Mack, and Roach Anleu (2018) have used an ethnographic approach to study the transcript and AV recordings in the High Court of Australia but the study is conducted in an observational manner via matching the Justices' distinct behaviour with the transcript.

1.2.2 Facial recognition

An anatomical analysis of facial action (Ekman and Friesen, 1976) led to the Facial Action Code (FAC) (Ekman and riesen, 1978) and has been further revised by Ekman, Friesen, and Hager (2002). This decomposition of facial muscles is widely used in scientific research. It was applied in competitive sports, specifically tennis by Kovalchik and Reid (2018) who found that the emotion of professional tennis players will have an impact on their performance.

There have been many algorithms created for facial detection and the analysis of their performance when applied to images have been the focus of events like the Audio/Visual Emotion Challenge (Schuller et al., 2012, 2011) and Emotion Recognition In The Wild Challenge and Workshop (Dhall et al., 2013; Kahou et al., 2013).

Facial recognition software has also been implemented by DeepFace (Taigman et al., 2014) from Facebook, and FaceNet (Schroff, Kalenichenko, and Philbin, 2015) from Google.

OpenFace (Baltrusaitis et al., 2018) is the first open-sourced face recognition software that provides facial expression detection, including facial landmarking, head pose estimation, eye gaze tracking and facial action unit detection. Along with its previous version (Baltrušaitis, Robinson, and Morency (2016)), the OpenFace toolkit has been used in different social research studies including depression classification (Yang et al. (2016) and Nasir et al. (2016)).

Chapter 2

Data Collection

2.1 Data Processing

The source data for this research project is the AV recordings publicly available from the High Court of Australia (Australia, 2019). Multiple procedures need to be performed to obtain the dataset.

The workflow for extracting numerical data from the videos can be found in Figure 2.1. Youtube-dl (Hsuan, Amine, and Sergey, 2019) has been used to download videos from the High Court of Australia(Australia, 2019) and the list of videos used in this research project is documented in the Appendix. Image frames are extracted from the videos for every minute via ffmpeg (Bellard, 2019), resulting in 1021 image frames (252 frames from `Nauru` videos and 769 frames from other five videos). Taipan (Kobakian and O'Hara-Wild, 2018) is then used to find the x-y coordinates of the location of the Justices in each image frame. ImageMagick (Cristy et al., 2019) is then used to crop the face of each Justice from each image frame that is taken from each video where three Justices present in `Nauru` videos and five Justices in other videos. The resulting 4601 cropped images (252 image frames from `Nauru` videos where three Justices are presented and 769 image frames from other videos where five justices are presented) are then sent to OpenFace (Baltrusaitis et al., 2018) to produce the variables for facial landmarking, head pose, eye gaze and facial action unit. This step is performed via the docker platform. The resulting outputs from

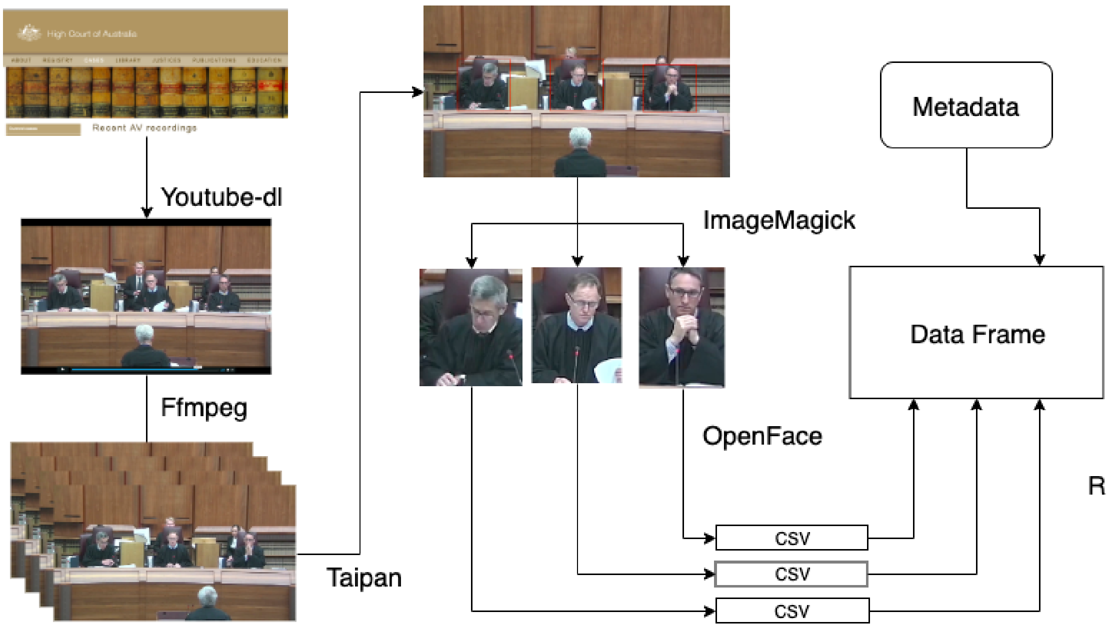


Figure 2.1: *data processing workflow*

OpenFace are individual comma-separated values (csv) files for each of the 4601 faces considered and processing is done in R to combine all the separate csv files into a final dataframe with appropriate index of `frame_id`, `judge_id` and `video_id`.

2.2 Variable description

OpenFace provides more than 711 variables measuring different aspect of a given face and a full description of the output variables can be found here. This outlines the difficulty of this project: no existing models will present accurate prediction and inference using 700+ variables - how can we incorporate these information to say about the facial expressions of the Justices during the hearings?

I conduct some exploratory data analysis on one video: `Nauru_a` and find the 700+ variables can be classified as follows with some insights

- **Confidence**: How confidence OpenFace is with the detection. Confidence is related to the angle that the Justice's face present in the images.
- **Gaze**: Gaze tracking: the vector from the pupil to corneal reflection. The dataset contains information on the gaze for both eyes while there is no distinct difference

between the eyes. Also I was trying to make animation to track the change of the gaze for judges but no good luck.

- **Pose**: the location of the head with respect to camera. Pose-related variables don't provide much useful information apart from gaze-related variables.
- **Landmarking**: landmarking variables for face and eyes. Landmarking variables allows me to plot the face of the judge in a particular frame. More work could be done to explore the usefulness of landmarking variables.
- **Action Unit**: Action units are used to describe facial expressions. this website provides a good animation on each action unit. The action unit has intensity measures ending with `_c` and presence measures ending with `_r`. These variables will be the focus of my project and a reference study of using action units to detect human emotion by Kovalchik can be found here.

2.3 Data format

In this project, we will make use of the action unit variables along with all the added indexes to analyse the face of the judge. In the wide format, apart from the first four index columns, each action unit has two columns with one for binary presence value and another for numeric intensity value. The Table 2.1 presents the first five rows of the dataset with columns for the first action unit only.

Table 2.1: *data in wide format*

judge	video	frame	AU01-r	AU02-r	AU04-r	AU05-r	AU06-r	AU07-r	AU09-r	AU10-
Bell	McKell	1	0	0	0.69	0.63	0	1.5	0	0
Bell	McKell	2	0	0	0.69	0.63	0	1.5	0	0

The data can also be expressed in the long format with action unit being another index and presence and intensity being two columns. The Table 2.2 presents the first five rows of the data in the long format.

Table 2.2: *data in long format*

judge	video	frame	speaker	AU	presence	intensity
Bell	McKell	1	Appellent	AU01	1	0.00
Bell	McKell	1	Appellent	AU02	1	0.00
Bell	McKell	1	Appellent	AU04	0	0.69
Bell	McKell	1	Appellent	AU05	1	0.63
Bell	McKell	1	Appellent	AU06	0	0.00
Bell	McKell	1	Appellent	AU07	1	1.54
Bell	McKell	1	Appellent	AU09	1	0.00
Bell	McKell	1	Appellent	AU10	1	0.00
Bell	McKell	1	Appellent	AU12	0	0.00
Bell	McKell	1	Appellent	AU14	0	0.00
Bell	McKell	1	Appellent	AU15	1	0.00
Bell	McKell	1	Appellent	AU17	0	0.00
Bell	McKell	1	Appellent	AU20	1	0.05
Bell	McKell	1	Appellent	AU23	0	0.00
Bell	McKell	1	Appellent	AU25	1	0.00
Bell	McKell	1	Appellent	AU26	0	0.26
Bell	McKell	1	Appellent	AU28	NA	NA
Bell	McKell	1	Appellent	AU45	0	0.47
Bell	McKell	2	Appellent	AU01	0	0.00
Bell	McKell	2	Appellent	AU02	1	0.00
Bell	McKell	2	Appellent	AU04	1	0.69
Bell	McKell	2	Appellent	AU05	1	0.63
Bell	McKell	2	Appellent	AU06	0	0.00
Bell	McKell	2	Appellent	AU07	1	1.54
Bell	McKell	2	Appellent	AU09	1	0.00
Bell	McKell	2	Appellent	AU10	1	0.00
Bell	McKell	2	Appellent	AU12	0	0.00
Bell	McKell	2	Appellent	AU14	0	0.00
Bell	McKell	2	Appellent	AU15	1	0.00
Bell	McKell	2	Appellent	AU17	0	0.00
Bell	McKell	2	Appellent	AU20	1	0.05
Bell	McKell	2	Appellent	AU23	0	0.00
Bell	McKell	2	Appellent	AU25	1	0.00
Bell	McKell	2	Appellent	AU26	0	0.26
Bell	McKell	2	Appellent	AU28	NA	NA
Bell	McKell	2	Appellent	AU45	0	0.47

2.4 Missing value imputation

The missingness in the dataset could be due to the fact that a judge is reading the materials on the desk so the face is not captured for a particular frame or simply because some faces are not detectable for the given resolution of the video stream. However, since that data is in time series structure, simply drop the missing observation will cause the time interval to be irregular and complicate further analysis.

There are two different sets of variables that need imputation. `Presence` is a binary variable that takes value of one if an action unit is present in a particular frame for a judge in a video and `Intensity` measures how strong that action unit is. Linear interpolation from `forecast` package is suitable to impute `Intensity` and `Presence` is imputed through sampling from binomial distribution. The imputed action unit data is stored as `au_imputed` under the `raw_data` folder.

2.5 Data cleaning

There is a data quality issue coming from the data I get from OpenFace. For some observations, the intensity of the action unit could be high while the present variable has a zero value. This does not make sense since if an action unit has been detected as strong intensity for a judge in a particular frame, it should at least present on the judge's face. Therefore, I adjust for the presence value if the intensity is higher than one. One is being chosen as the threshold value since in Ekman's definition of the intensity of the action unit, a score of one means the action unit is at least slightly present in the judge's face. The adjusted data is stored as `au_tidy` under the `raw_data` folder.

Chapter 3

Methods

Here you need to write about the analytical methods that you are using

Chapter 4

Results

4.1 Notation

Let $\mathbf{X}$ be a matrix of predictors, and $\mathbf{Y}$ variable in our case is bivariate matrix of response variables, including a binary indicator of presence/absence and a numeric value measuring intensity, of facial action unit, where

- X_1 indicates `judge` with six categories $i = 1, 2, \cdots, 6$
- X_2 indicates `video` for each of the seven cases, $j = 1, 2, \cdots, 7$
- X_3 indicates action unit containing 18 possible facial expression.

- X_4 indicates `speaker`, either the appellant or respondent, $l = 1, 2$
- X_5 indicates `frame` corresponding to time, $t = 1, 2, \cdots, T_j$

Note that t could be considered a time variable, but because images are taken at 1 minute intervals, temporal dependence is unlikely to exist. Rather this should be considered an independent observation.

A full, main effects model for the data might be expressed as:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \varepsilon_{ijkl}$$

and we would be interested in interactions between judge, case, action unit and who is speaking. An alternative model structure, is to treat each action unit individually, and fit separate models.

Also, let P_{jitkl} represent the response variable presence, and I_{jitkl} represent the second response variable intensity. This notation will be helpful for defining the plots and models explained in this section.

4.2 Action unit: Presence

4.2.1 Mean presence

I first compute the average presence (P_{ik}) of each action unit for each judge as

$$P_{ik} = \frac{\sum_{jt} X_{ijtk}}{\sum_{j=1}^{J} T_j}$$

This is then plotted in Figure 4.1 to give an overview of the presence of all the action units across all the judge. The order of action unit on the y axis is ranked by the average presence of all the judge. The five most frequent action units are highlighted in blue for each judge and summarised in Table 4.1

It can be seen that some of the action units are common across almost all the judges, these includes

- AU02 (outer eyebrow raise),
- AU20 (lip stretcher),
- AU15 (Lip Corner Depressor)
- AU14 (Dimpler)

AU02 makes a contribution to surprise, which is a positive attitude showing that judges are interested in a particular moment (Ekman, Friesen, and Hager, 2002). According to (Ekman, Friesen, and Hager, 2002), AU14 indicates boredom and AU15 shows confusion. Along with other action units that presented with high frequency in a particular judge but not all (summarised in Table 4.3), the emotions judges displayed in the courtroom can be

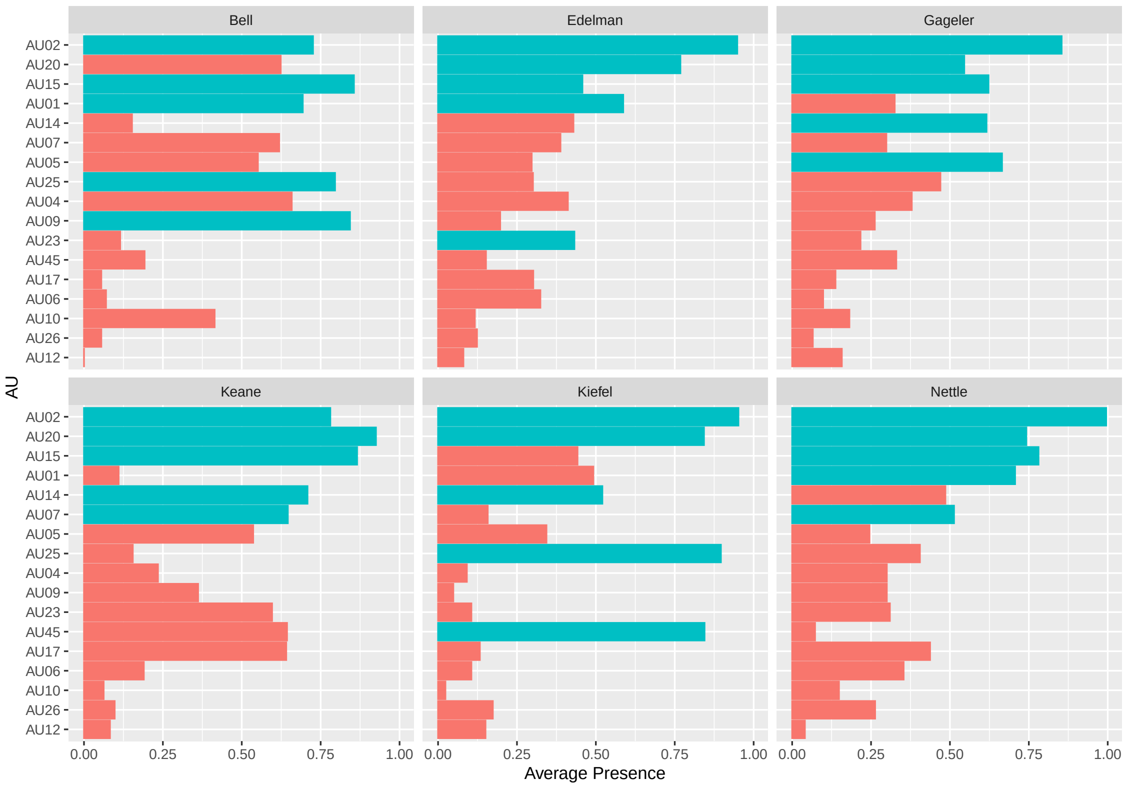


Figure 4.1: *The average presence score of each action unit for each judge, aggregating on video and time.*

Table 4.1: *The five most commonly presented action unit for each judge.*

index	Bell	Edelman	Gageler	Keane	Kiefel	Nettle
1	AU15	AU02	AU02	AU20	AU02	AU02
2	AU09	AU20	AU05	AU15	AU25	AU15
3	AU25	AU01	AU15	AU02	AU45	AU20
4	AU02	AU15	AU14	AU14	AU20	AU01
5	AU01	AU23	AU20	AU07	AU14	AU07

summarised into three categories, described in Table 4.2 along with the featured action units.

4.2.2 Model fit

The first model I use is a generalised linear model with binomial link to understand the presence of the action units. The variables used include the judge, action units and their interactions. The model can be written down as Equation 4.1.Judge Edelman and AU01 are selected as the base level.

Table 4.2: *Summarised emotions and featured action units*

emotion	Featured Action Unit
Surprise	AU01, AU02, AU05
Boredom	AU14, AU23
Confusion	AU07, AU15, AU23

Table 4.3: *Other high frequent action units*

judge	first	second	third
Bell	AU09: Nose wrinkler	AU25: Lips part	AU01: Inner brow raiser
Edelman	AU01: Inner brow raiser	AU23: Lip tightener	NA
Gageler	AU05: Upper lid raiser	NA	NA
Keane	AU07: Lid tightener	NA	NA
Kiefel	AU25: Lips part	AU45: Blink	NA
Nettle	AU01: Inner brow raiser	AU07: Lid tightener	NA

$$P_{ik} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} \tag{4.1}$$

I'm interested to know if the presence score for one pair of judge and action unit is signficiantly different from another pair. Aanalysis of Varaince (ANOVA) test result in Table 4.4 shows that there are significant variance for both judge, AU and their interactions. The next step after ANOVA is to test how each level of judge and AU different from another and I use multiple comparison to do this. (manually it will be 861 test to perform since 6 judges and 7 au - using multiple comparison, we can perform this take while control for a relatively low 5% false positive rate). The estimated coefficient for each judge and AU pair is reported in Table 4.5 along with the upper and lower confidence interval bond. The information in the group column is helpful to understand how one particular pair of judge and au is different from another pair. With compact letter display, the pair with the same letter/number are *NOT* significantly different from each other. This result is also plotted in Figure 4.2.

Table 4.4: *ANOVA result*

	Df	Deviance	Resid. Df	Resid. Dev
NULL	NA	NA	15183	18900.38
judge	5	298.5383	15178	18601.84
AU	3	1682.9621	15175	16918.88
judge:AU	15	1113.8722	15160	15805.01

Table 4.5: *model result*

judge	AU	prob	SE	df	asymp.LCL	asymp.UCL	.group
Bell	AU14	0.15	0.0179	Inf	0.12	0.19	1
Edelman	AU14	0.43	0.0155	Inf	0.40	0.46	2
Kiefel	AU15	0.44	0.0210	Inf	0.40	0.48	2
Edelman	AU15	0.46	0.0156	Inf	0.43	0.49	2
Nettle	AU14	0.49	0.0201	Inf	0.45	0.52	23
Kiefel	AU14	0.52	0.0211	Inf	0.48	0.56	234
Gageler	AU20	0.55	0.0177	Inf	0.51	0.58	345
Gageler	AU14	0.62	0.0173	Inf	0.58	0.65	456
Gageler	AU15	0.62	0.0172	Inf	0.59	0.66	567
Bell	AU20	0.62	0.0242	Inf	0.57	0.67	4567
Keane	AU14	0.71	0.0227	Inf	0.66	0.75	678
Bell	AU02	0.73	0.0223	Inf	0.68	0.77	78
Nettle	AU20	0.74	0.0176	Inf	0.71	0.77	8
Edelman	AU20	0.77	0.0132	Inf	0.74	0.79	89
Keane	AU02	0.78	0.0207	Inf	0.74	0.82	890
Nettle	AU15	0.78	0.0166	Inf	0.75	0.81	890
Kiefel	AU20	0.84	0.0154	Inf	0.81	0.87	90
Gageler	AU02	0.85	0.0125	Inf	0.83	0.88	0A
Bell	AU15	0.86	0.0176	Inf	0.82	0.89	90A
Keane	AU15	0.87	0.0170	Inf	0.83	0.90	0A
Keane	AU20	0.93	0.0131	Inf	0.90	0.95	AB
Edelman	AU02	0.95	0.0069	Inf	0.93	0.96	B
Kiefel	AU02	0.95	0.0091	Inf	0.93	0.97	B
Nettle	AU02	1.00	0.0028	Inf	0.99	1.00	C

4.2.3 Presence by videos

Apart from visualising the general presence score for all the action units, I'm also interested in the break down statistics by video (P_{ijk}). This is computed as

$$P_{ijk} = \frac{\sum_t X_{ijtk}}{T_j}$$

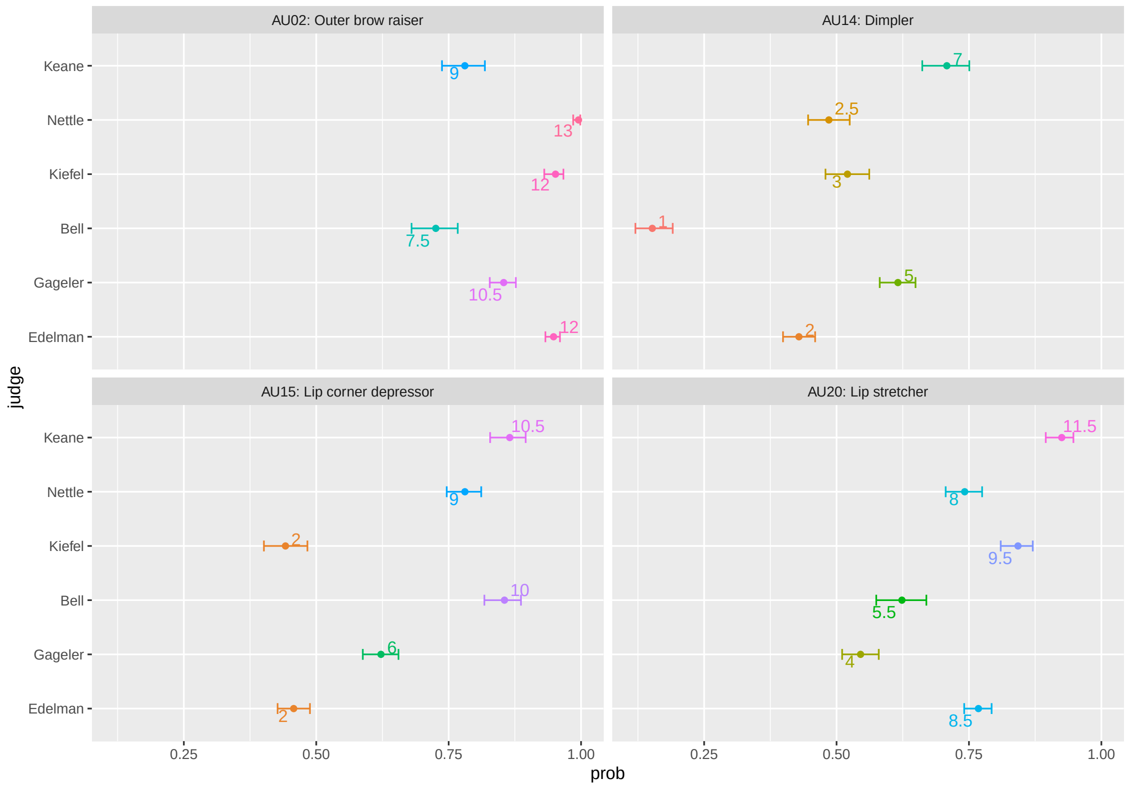


Figure 4.2: *THis is the graphical representation of model1*

for the four most common action units: AU02, AU14, AU15, AU20 and plotted in Figure 4.3. From this plot, we can observe that some of the judge are have relatively stable display of action unit throughout different videos (i.e. Edelman and Nettle), while Gagaler seems to be highly reactive to some cases (i.e. OKS).

4.2.4 Model fit

The second model as shown in Equation 4.2 is estimated to understand the interaction effect between judge and video.

$$P_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + (\alpha\beta)_{ij} \tag{4.2}$$

The estimated coefficients are presented in Table 4.6

What we could find from Figure 4.4

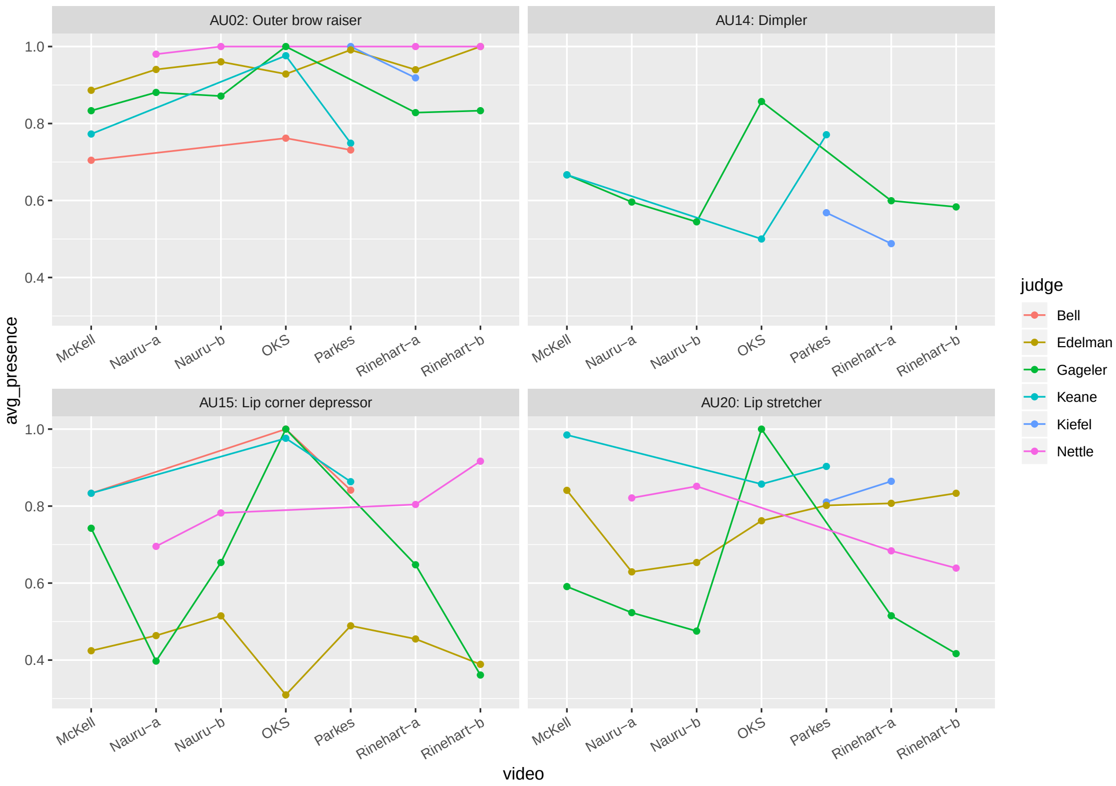


Figure 4.3: *Average presence of the four most common action units for each judge by video*

- Judge Edelman, Keane and Kiefel behave relatively consistent throughout all the videos. Judge Gageler is also consistent throughout the trails except for video OKS.
- Judge Nettle seems to have two different "status"
- Judge Bell behaves quite differently in the three videos she participates

4.2.5 Appellant vs. Respondent

The third model as shown in Equation 4.3 is estimated to understand the interaction effect between judge and speaking party.

$$P_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\delta)_{il} \tag{4.3}$$

The estimated coefficients are presented in Table 4.7

What we could find from Figure 4.5

- Judges are behaving pretty similar when different parties are talking

Table 4.6: *model result 2*

judge	video	prob	SE	df	asymp.LCL	asymp.UCL	.group
Bell	Parkes	0.50	0.018	Inf	0.47	0.54	1
Gageler	Rinehart-b	0.56	0.044	Inf	0.47	0.65	123
Edelman	OKS	0.58	0.041	Inf	0.50	0.66	1234
Gageler	Nauru-a	0.62	0.021	Inf	0.58	0.66	2
Gageler	Nauru-b	0.66	0.025	Inf	0.61	0.71	234
Edelman	Nauru-a	0.67	0.020	Inf	0.63	0.71	234
Gageler	Rinehart-a	0.67	0.014	Inf	0.65	0.70	234
Bell	McKell	0.68	0.021	Inf	0.64	0.72	234
Edelman	McKell	0.68	0.021	Inf	0.64	0.72	234
Edelman	Parkes	0.68	0.016	Inf	0.65	0.71	234
Edelman	Nauru-b	0.68	0.024	Inf	0.63	0.73	2345
Edelman	Rinehart-a	0.68	0.013	Inf	0.66	0.71	234
Edelman	Rinehart-b	0.69	0.040	Inf	0.61	0.77	23456
Kiefel	Rinehart-a	0.71	0.013	Inf	0.68	0.73	2345
Kiefel	Parkes	0.74	0.015	Inf	0.70	0.76	456
Gageler	McKell	0.74	0.020	Inf	0.70	0.77	3456
Nettle	Nauru-a	0.74	0.018	Inf	0.71	0.78	456
Nettle	Rinehart-a	0.77	0.012	Inf	0.75	0.79	567
Nettle	Nauru-b	0.83	0.019	Inf	0.79	0.86	678
Keane	McKell	0.84	0.015	Inf	0.81	0.87	78
Keane	Parkes	0.85	0.012	Inf	0.83	0.87	8
Keane	OKS	0.86	0.026	Inf	0.80	0.90	6789
Nettle	Rinehart-b	0.88	0.026	Inf	0.82	0.92	6789
Bell	OKS	0.90	0.022	Inf	0.85	0.93	789
Gageler	OKS	0.97	0.011	Inf	0.94	0.99	9

Table 4.7: *model result 3*

judge	speaker	prob	SE	df	asymp.LCL	asymp.UCL	.group
Bell	Appellent	0.60	0.021	Inf	0.56	0.64	1
Bell	Respondent	0.61	0.020	Inf	0.57	0.65	1
Edelman	Appellent	0.69	0.011	Inf	0.67	0.71	2
Gageler	Appellent	0.70	0.012	Inf	0.67	0.72	2
Edelman	Respondent	0.71	0.012	Inf	0.68	0.73	2
Gageler	Respondent	0.72	0.013	Inf	0.69	0.74	2
Kiefel	Respondent	0.73	0.014	Inf	0.70	0.75	2
Kiefel	Appellent	0.79	0.015	Inf	0.76	0.82	3
Nettle	Respondent	0.80	0.012	Inf	0.77	0.82	3
Nettle	Appellent	0.82	0.011	Inf	0.80	0.84	34
Keane	Appellent	0.83	0.015	Inf	0.80	0.86	34
Keane	Respondent	0.87	0.012	Inf	0.84	0.89	4

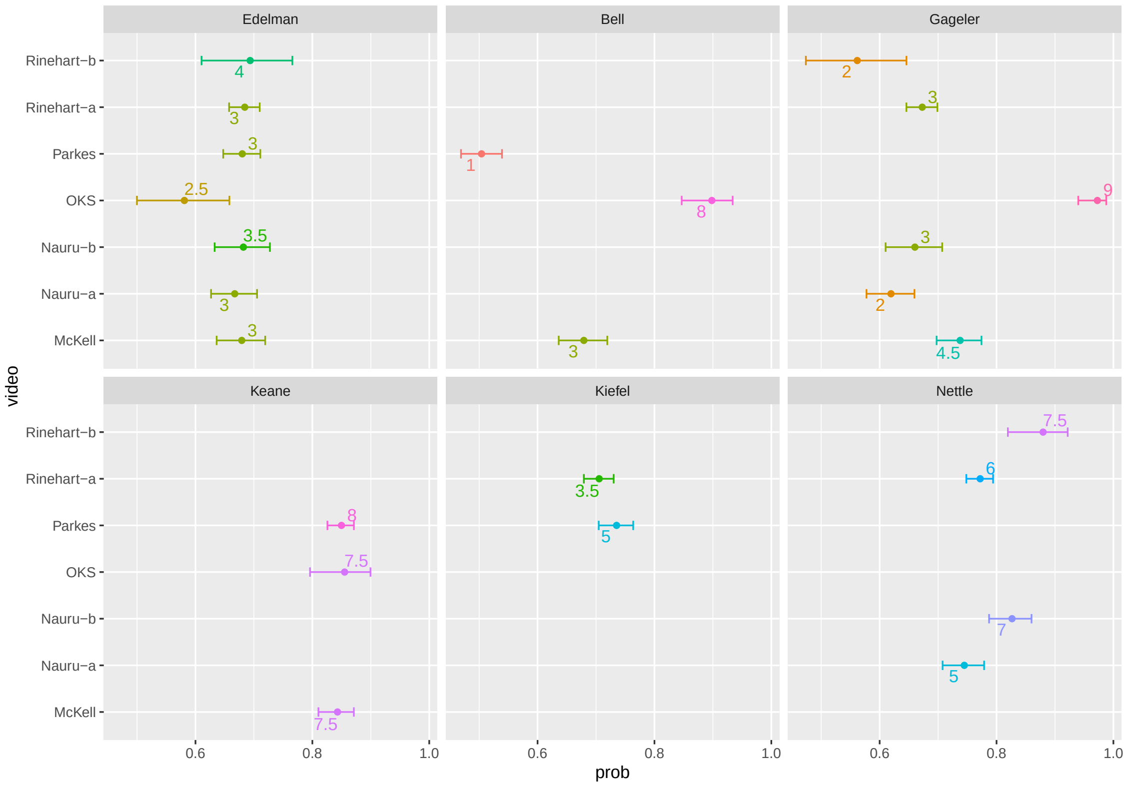


Figure 4.4: *THis is the graphical representation of model2*

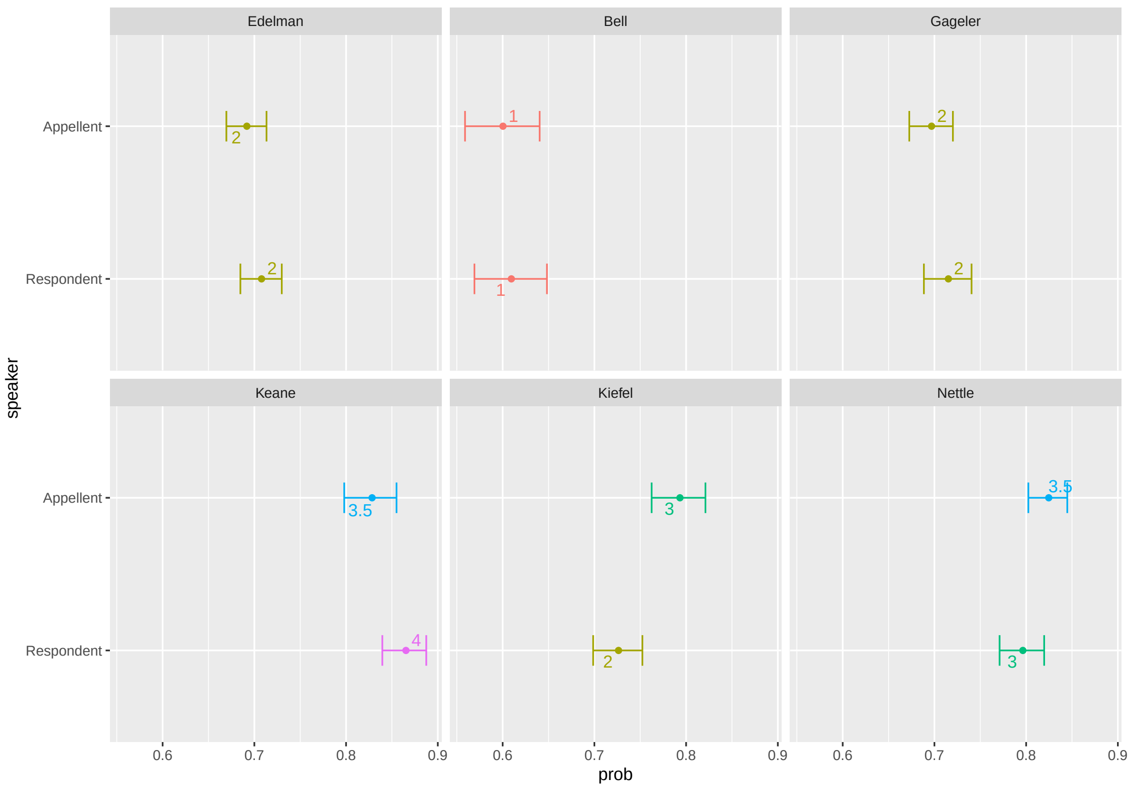


Figure 4.5: *THis is the graphical representation of model3*

4.3 Action unit: Intensity

4.3.1 General Intensity plot

In Ekman's 20002 FACS manual, the intensity of an action unit is defined based on five classes: Trace: 0-1, Slight: 1-2, Marked or pronounced: 2-3, Severe or extreme: 3-4 and Maximum: 4-5.

The boxplot of the intensity for all the judges across all the videos is presented in Figure 4.6. Each bar-and-whisker represents the intensity (I_{ijtk}) of all the action units aggregated on time for a particular judge i in a specific case j. For example, the first bar-and-whisker in case Nauru_a is created using all the 17 action units of Edelman through out the elapsed time in Nauru_a case.

From the plot, we can see that most of the action units have low intensity score and this is expected because usually judges are expected to behave neutral in the court room. Thus a square root transformation is taken on the y axis for better visualisation effect. We can find that Judge Nettle seems to have higher average in all the four cases he appears: Nauru_a&b, Rinehart_a &b.

4.3.2 Mean intensity

Mean intensity score (I_{ik}) of each action unit for each of the judge is computed as

$$I_{ik} = \frac{\sum_{jt} X_{ijtk}}{\sum_{j=1}^{J} T_j}$$

and plotted in Figure 4.7. The five most intense action units for each judge are presented in Table 4.8. We can find that the common high intense action units includes

- AU20 (Lip Stretcher)
- AU07 (Lid Tightener)
- AU04 (Brow Lowerer)

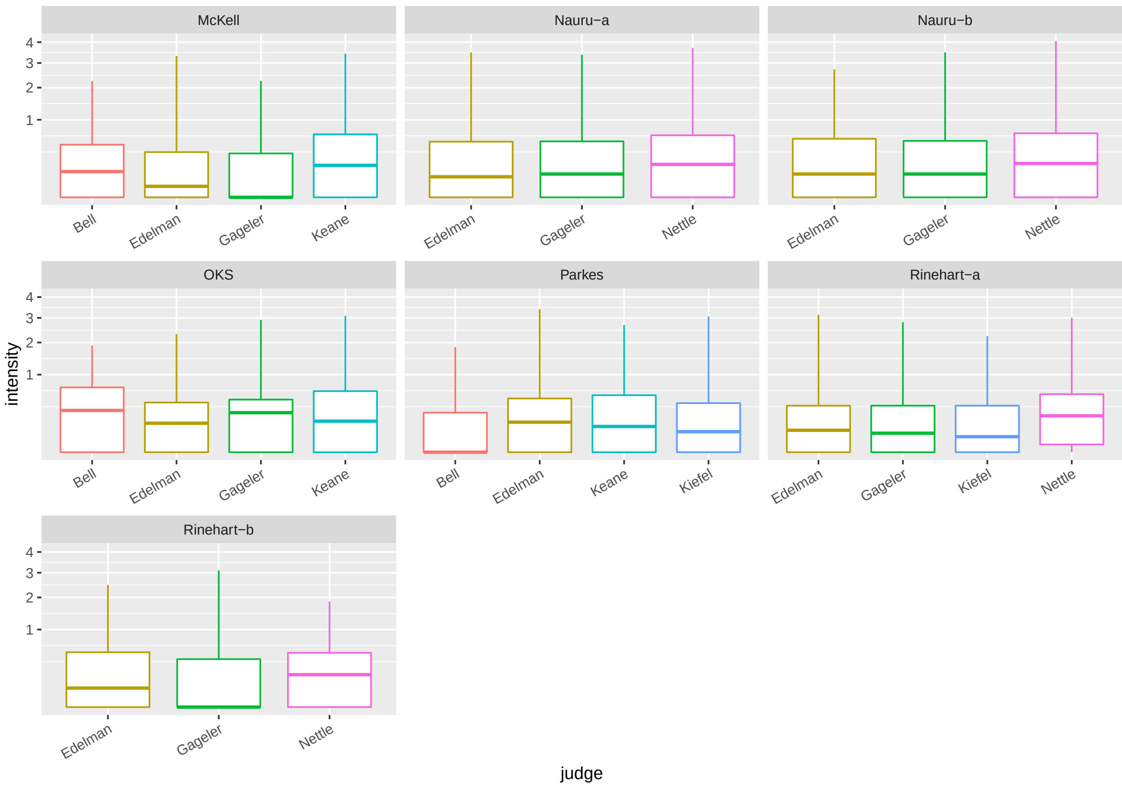


Figure 4.6: *General intensity score by judge and video*

Table 4.8: *The five most intense action unit for each judge.*

index	Bell	Edelman	Gageler	Keane	Kiefel	Nettle
1	AU04	AU20	AU05	AU20	AU45	AU20
2	AU09	AU06	AU04	AU07	AU20	AU01
3	AU07	AU17	AU20	AU15	AU26	AU07
4	AU15	AU07	AU09	AU14	AU25	AU06
5	AU10	AU04	AU45	AU17	AU14	AU17

AU04 also belongs to the confusion category as AU07. This could help to understand that judges are more likely to express a stronger confusing expression than other emotions.

4.3.3 Model fit

4.3.4 Intensity plot for the most frequent action units

Apart from visualising the general intensity score for all the action units, I'm also interested in the intensity score of the most frequent action units. Figure 4.8 presents this. The statistics being plotted is I_{ijtk} with k including AU02, AU14, AU15 and AU20 as the

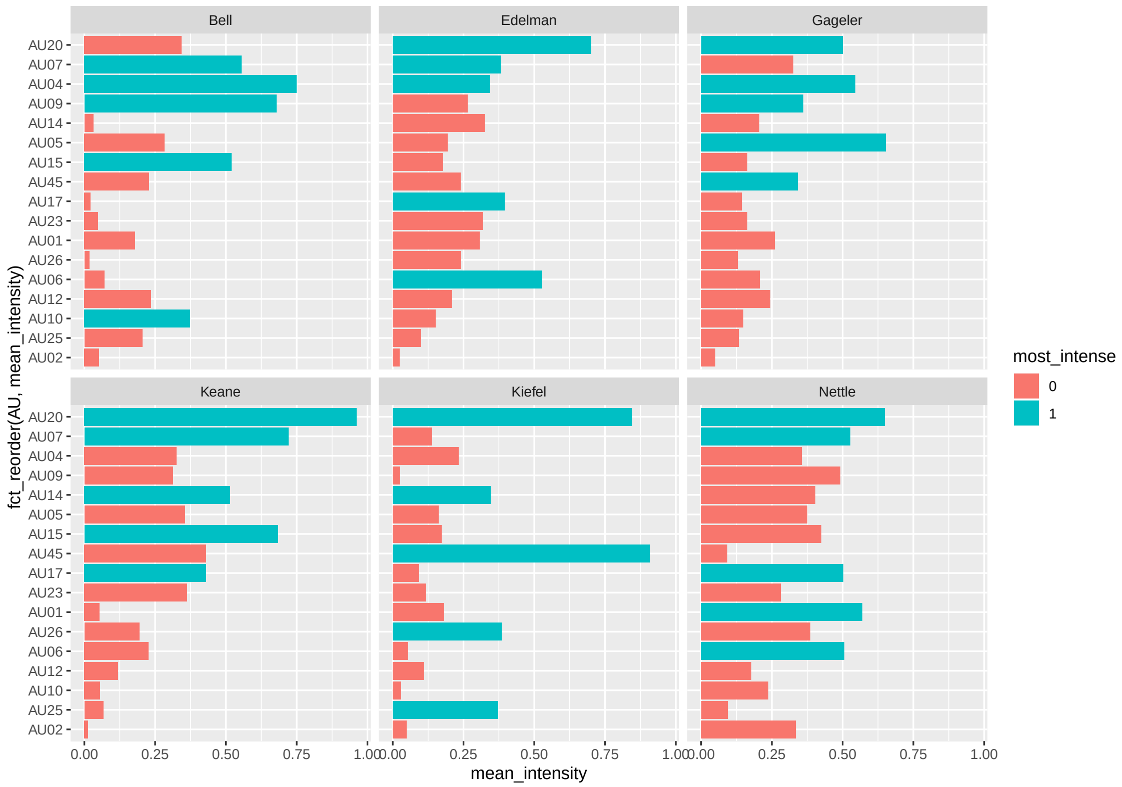


Figure 4.7: *Mean intensity score for each judge and action unit aggregating on videos.*

most common four action units. From this plot, we can learn that AU02, although being commonly detected for all the judges, has low intensity score.

4.3.5 High intensity points

We filter out the points have intensity greater than 2 (at least "slight" as per Ekman) in the previous plot and plot it against time and color by the speaker. It tells us that Edelman, Gageler and Nettle are the judges have stronger emotion that can be detected (since they have more points with intensity greater than 2). Different judges also have different time where they display stronger emotions. For example, Justice Nettle are more likely to have stronger emotion throughout the time when the appellant is speaking but only at the beginning and ending period when the respondent is speaking.

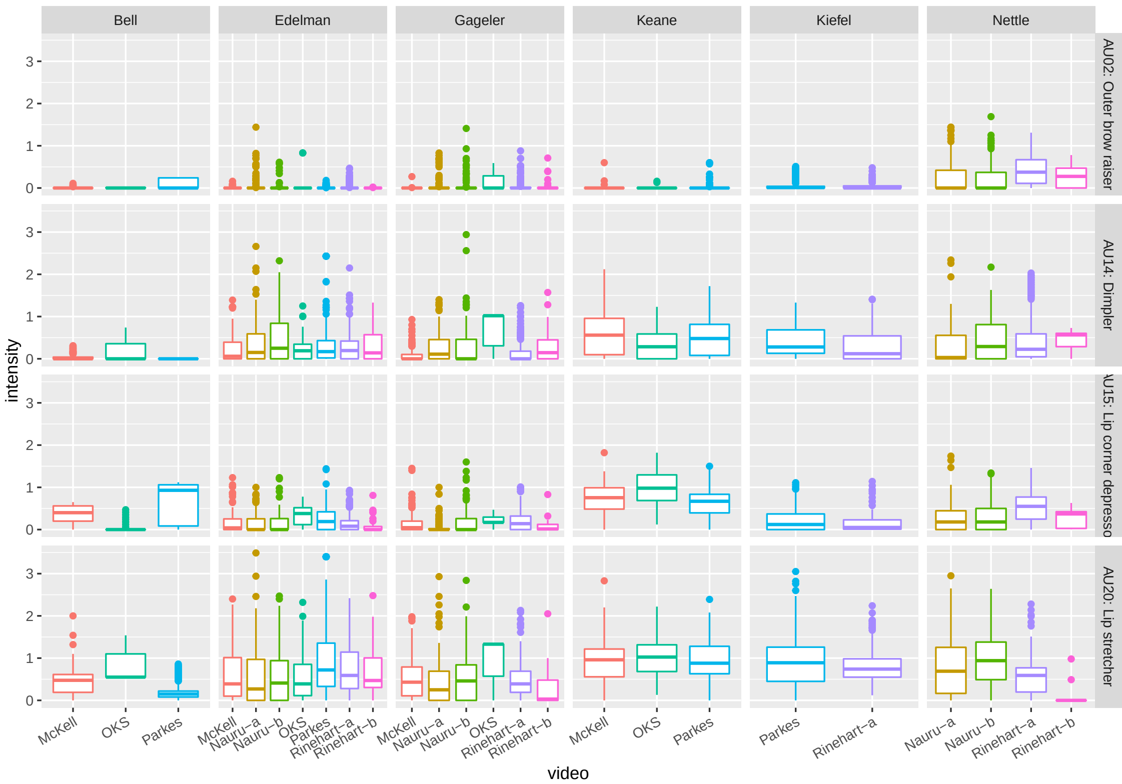
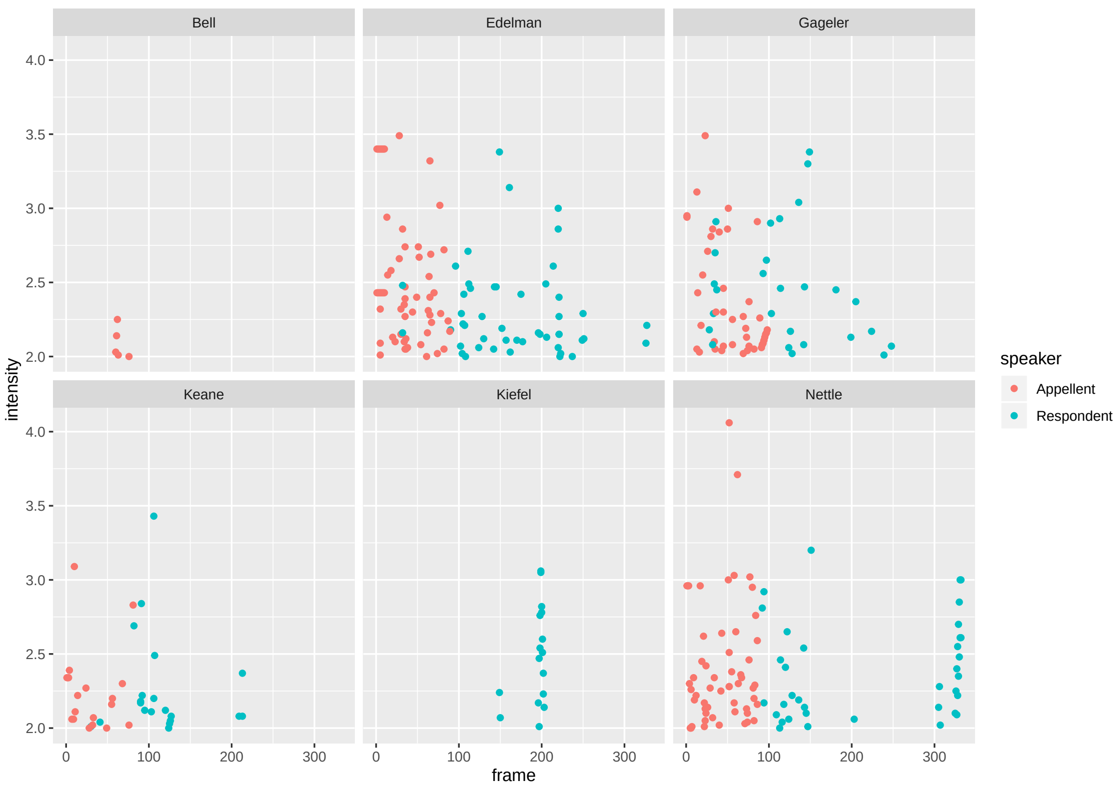


Figure 4.8: *Intensity score of the most frequent action units, seperating by judge and video ID.*



Appendix A

Additional stuff

Table A.1: *Details of videos processed.*

Case	Name	AV recording link
Republic of Nauru v. WET040	Nauru_a	http://www.hcourt.gov.au/cases/cases-av/av-2018-11-07a
TTY167 v. Republic of Nauru	Nauru_b	http://www.hcourt.gov.au/cases/cases-av/av-2018-11-07b
Rinehart & Anor v. Hancock Prospecting Pty Ltd & Ors on 13 Nov 18	Rinehart_a	http://www.hcourt.gov.au/cases/cases-av/av-2018-11-13
Rinehart & Anor v. Hancock Prospecting Pty Ltd & Ors on 14 Nov 18	Rinehart_b	http://www.hcourt.gov.au/cases/cases-av/av-2018-11-14a
Parkes Shire Council v. South West Helicopters Pty Limited	Parkes	http://www.hcourt.gov.au/cases/cases-av/av-2018-11-14b

Case	Name	AV recording link
McKell v. The Queen	McKell	http://www.hcourt.gov.au/cases/cases-av/av-2018-12-07
OKS v. The State of Western Australia	OKS	http://www.hcourt.gov.au/cases/cases-av/av-2019-02-14

A.1 List of the name of ction units

AU_meaning
AU01: Inner brow raiser
AU02: Outer brow raiser
AU04: Brow lowerer
AU05: Upper lid raiser
AU06: Cheek raiser
AU07: Lid tightener
AU09: Nose wrinkler
AU10: Upper lip raiser
AU12: Lip corner puller
AU14: Dimpler
AU15: Lip corner depressor
AU17: Chin raiser
AU20: Lip stretcher
AU23: Lip tightener
AU25: Lips part
AU26: Jaw drop
AU28: Lip suck
AU45: Blink

Bibliography

Australia, HC of (2019). *Recent AV recordings*. Accessed: 2019-05-03. http://www.hcourt.gov.au/cases/recent-av-recordings.

Baltrušaitis, T, P Robinson, and LP Morency (2016). Openface: an open source facial behavior analysis toolkit. In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp.1–10.

Baltrusaitis, T, A Zadeh, YC Lim, and LP Morency (2018). Openface 2.0: Facial behavior analysis toolkit. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pp.59–66.

Bellard, F (2019). *ffmpeg*. https://ffmpeg.org/.

Black, RC, SA Treul, TR Johnson, and J Goldman (2011). Emotions, oral arguments, and Supreme Court decision making. *The Journal of Politics* **73**(2), 572–581.

Chen, DL, M Kumar, V Motwani, and P Yeres (2018). *Is Justice Really Blind? And Is It Also Deaf*. Tech. rep. Technical report.

Chen, D, Y Halberstam, and C Alan (2016). Perceived masculinity predicts us supreme court outcomes. *PloS one* **11**(10), e0164324.

Chen, D, Y Halberstam, A Yu, et al. (2017). Covering: Mutable characteristics and perceptions of voice in the US Supreme Court. *Review of Economic Studies invited to resubmit, TSE Working Paper* (16-680).

Chief Justices of Australia, TC of and N Zealand (2017). *Guide to Judicial Conduct*. 3rd. Melbourne: Austral-asian Institute of Judicial Administration.

Cristy, J, D Lemstra, G Randers-Pehrson, and B Roucries (2019). *ImageMagick-dl*. https://github.com/ImageMagick.

Dhall, A, R Goecke, J Joshi, M Wagner, and T Gedeon (2013). Emotion recognition in the wild challenge 2013. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, pp.509–516.

Dietrich, BJ, RD Enos, and M Sen (2019). Emotional arousal predicts voting on the US supreme court. *Political Analysis* **27**(2), 237–243.

Ekman, P and WV Friesen (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior* **1**(1), 56–75.

Ekman, P, WV Friesen, and JC Hager (2002). Facial action coding system: The manual on CD ROM. *A Human Face, Salt Lake City*, 77–254.

Ekman, P, M O'Sullivan, WV Friesen, and KR Scherer (1991). Invited article: Face, voice, and body in detecting deceit. *Journal of nonverbal behavior* **15**(2), 125–135.

Ekman, P and WV riesen (1978). *Facial action coding system*. Palo Alto: CA: Consulting Psychologists Press.

Epstein, L, WM Landes, and RA Posner (2010). Inferring the winning party in the Supreme Court from the pattern of questioning at oral argument. *The Journal of Legal Studies* **39**(2), 433–467.

Goffman, E (1956). The nature of deference and demeanor. *American Anthropologist* **58**(3), 473–502.

Hsuan, YC, R Amine, and M Sergey (2019). *youtube-dl*. https://github.com/ytdl-org/youtube-dl/.

Johnson, TR, RC Black, J Goldman, and SA Treul (2009). Inquiring minds want to know: Do justices tip their hands with questions at oral argument in the US supreme court. *Washington University Journal of Law and Policy* **29**, 241.

Kahou, SE, C Pal, X Bouthillier, P Froumenty, Ç Gülçehre, R Memisevic, P Vincent, A Courville, Y Bengio, RC Ferrari, et al. (2013). Combining modality specific deep neural networks for emotion recognition in video. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, pp.543–550.

Kobakian, S and M O'Hara-Wild (2018). *taipan: Tool for Annotating Images in Preparation for Analysis*. R package version 0.1.2. https://CRAN.R-project.org/package=taipan.

Kovalchik, S and M Reid (2018). Going inside the inner game: Predicting the emotions of professional tennis players from match broadcasts. In: MIT Sloan Sports Analytics Conference. Accessible at http://www ...

Nasir, M, A Jati, PG Shivakumar, S Nallan Chakravarthula, and P Georgiou (2016). Multimodal and multiresolution depression detection from speech and facial landmark features. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pp.43–50.

Schroff, F, D Kalenichenko, and J Philbin (2015). Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.815–823.

Schubert, JN, SA Peterson, G Schubert, and S Wasby (1992). Observing Supreme Court oral argument: A biosocial approach. *Politics and the Life Sciences* **11**(1), 35–52.

Schuller, B, M Valstar, F Eyben, G McKeown, R Cowie, and M Pantic (2011). Avec 2011–the first international audio/visual emotion challenge. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp.415–424.

Schuller, B, M Valster, F Eyben, R Cowie, and M Pantic (2012). AVEC 2012: the continuous audio/visual emotion challenge. In: *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, pp.449–456.

Shullman, SL (2004). The illusion of devil's advocacy: How the justices of the supreme court foreshadow their decisions during oral argument. *Journal of Appellate Practice and Process 2* **271**(6).

Taigman, Y, M Yang, M Ranzato, and L Wolf (2014). Deepface: Closing the gap to human-level performance in face verification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.1701–1708.

Tutton, J, K Mack, and S Roach Anleu (2018). Judicial Demeanor: Oral Argument in the High Court of Australia. *Justice System Journal* **39**(3), 273–299.

Yang, L, D Jiang, L He, E Pei, MC Oveneke, and H Sahli (2016). Decision tree based depression classification from audio video and language information. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pp.89–96.