Measuring Audience and Actor Emotions at a Theater Play through Automatic Emotion Recognition from Face, Speech, and Body Sensors

Peter A. Gloor[1], **Keith April Araño**[2], **Emanuele Guerrazzi**[3]

[1] Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, USA. pgloor@mit.edu

[2] Politecnico di Milano, Via Lambruschini, 4/B, 20156, Milan, Italy

[3] University of Pisa, Largo Lucio Lazzarino, 56122, Pisa, Italy

Abstract We describe a preliminary experiment to track the emotions of actors and audience in a theater play through machine learning and AI. During a forty-minute play in Zurich, eight actors were equipped with body sensing smartwatches. At the same time, the emotions of the audience were tracked anonymously using facial emotion tracking. In parallel, also the emotions in the voices of the actors were assessed through automatic voice emotion tracking. This paper demonstrates a first fully automated and privacy-respecting system to measure both audience and actor satisfaction during a public performance.

1. Introduction

Emotion recognition has been widely studied for many years. Human emotion is a crucial element for communication and decision-making. The availability of emotion-rich data sources on many channels, along with recent advances in machine learning and deep learning have led to the development of various intelligent systems that are able to automatically recognize and interpret human emotions. In businesses for example, online retail systems are capable of analyzing emotional customer feedback to improve customer satisfaction (Hong, Zheng, Wu, & Pu, 2019). In healthcare, the physical and emotional states of patients are monitored to automatically diagnose and prescribe the appropriate treatment (Chen, Zhang, Qiu, Guizani, & Hao, 2018). Another application of emotion recognition is for safe driving through online monitoring of driver emotions (Vasey, Ko, & Jeon, 2018).

Traditionally, emotion recognition research has been focused on analyzing unimodal data: speech signals (Swain, Routray, & Kabisatpathy, 2018), text data (Yadollahi, Shahraki, & Zaiane, 2017), facial expressions (Ko, 2018), and most

recently, physiological signals (Ali, Mosa, Machot, & Kyamakya, 2018). However, emotions are complex cognitive processes with rich features that are difficult to infer with just a single modality (Qiu, Liu, & Lu, 2018). Consequently, a number of studies have investigated the use of multimodal data and have shown that it can substantially improve the prediction of emotional states (Ullah, Islam, Azman, & Zaki, 2017). Moreover, the concept of cross-modal prediction in which shared representations are learned from multiple modalities to predict emotion from one modality to another has recently received growing interest from the research community.

One of the most widely known challenges in the field of emotion recognition is the difficulty of obtaining and labelling datasets to train prediction models. Cross-modal prediction addresses this problem by learning embeddings from one modality to predict another. For example, Albanie and colleagues (Albanie, Nagrani, Vedaldi, & Zisserman, 2018) investigated the task of learning speech embeddings without access to any form of labelled audio data by exploiting a pre-trained face emotion recognition network, to reduce the dependence on labelled speech. Similarly, Li et. al (Li, Zhu, Tedrake, & Torralba, 2019) proposed a cross-modal prediction system between vision and touch that is capable of learning to see by touching, and learning to feel by seeing. Inter-modality dynamics, which models the interactions between different modalities and how they affect the expressed emotions of an individual, have also been investigated in earlier work (Zadeh, Chen, Poria, Cambria, & Morency, 2018). The majority of these studies on using multimodal data, however, have been focused on recognizing emotions of a single individual, and little research has explored such inter-modality interactions between a group of individuals.

Motivated by these advances in multimodal emotion recognition, we investigate the correlations between the emotions depicted from facial expressions, speech, and physiological signals between two separate groups of individuals. In particular, we monitor the interaction between actors and audience in a theatre performance. The contributions of this paper are as follows: We trained a face emotion recognition (FER) model and a speech emotion recognition (SER) model that are capable of predicting emotions from facial expressions and speech signals, respectively. We collected visual, audio, and physiological data from actors and audience of a theatre performance which took place at the Landesmuseum in Zurich, Switzerland in spring of 2019. To the best of our knowledge, no such dataset has been collected before. We finally investigated the correlations between the emotions predicted by our deep learning models from the facial expressions, speech, and physiological signals that we have collected. Specifically, we analyzed the emotions of actors from their speech and physiological signals, and how these emotions translate to the facial expressions of the audience, investigating inter-modal and inter-personal dynamics.

2. Theoretical Background

Psychologists have proposed several theories categorizing different emotions that also account for age and cultural differences. One of the most widely applied emotion categorization frameworks is Paul Ekman's emotion model (Ekman & Friesen, 1971) where he classifies emotions into six basic categories: anger, happiness, fear, surprise, disgust, and sadness. Another universally recognized emotion classification system is the Circumplex model of affect (Posner, Russell, & Peterson, 2005), which is a two-dimensional model with valence describing the range of negative and positive emotions, and arousal depicting the active to passive scale of emotions. High valence and high arousal for example, represent a pleasant feeling with high activation, which describes emotions such as happiness and excitement.

These emotions can be expressed in several ways: through facial expressions, speech, text, body language, or physiological signals. Among these modalities, facial expression is believed to be one of the most powerful and direct channels to convey human emotions in non-verbal communication (Ambady & Weisbuch, 2010; Rule & Ambady, 2010) while speech, on the other hand, is one of the most natural channels to transmit emotions in verbal interactions. These modalities differ in their potential in predicting emotional states as well as in their availability and usability under various circumstances (Egger, Ley, & Hanke, 2019). Moreover, one modality can be influential in the recognition of another, which has been investigated in prior studies in cross-modal prediction (Albanie et al., 2018; Li et al., 2019). This can be useful in applications where one modality is utilized when the other is absent, such as in generating captions or labels for images (Karpathy & Li, 2014) or in using vision to predict sounds (Owens et al., 2015).

Various methods have been proposed for recognizing emotions from faces, speech, and physiological signals. In face emotion recognition (FER), the current dominant technique are deep neural networks (DNNs) such as Convolutional Neural Networks (CNNs), which have been extensively used in diverse computer vision tasks that have resulted in several well-known CNN architectures such as AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGG (Simonyan & Zisserman, 2014), VFF-face (Parkhi, Vedaldi, & Zisserman, 2015), and GoogleNet (Szegedy et al., 2014). Similarly, in speech emotion recognition (SER), the recent breakthroughs in deep learning have led to the design of numerous DNN architectures such as variants of CNNs and Long-Short Term Memory (LSTM) networks, that have shown state-of-the art performance in SER (Lee & Tashev, 2015; Trigeorgis et al., 2016). In emotion recognition from physiological signals however, the majority of prior studies use classical algorithms such as Support Vector Machines (SVM), Random Forests (RF), Linear Discriminant Analysis (LDA), and K-nearest neighbors (KNN) (Ali et al., 2018). Deep learning in this domain is still in its infancy, possibly due to the lack of large physiological emotion-labeled datasets

necessary for training deep networks, contrary to FER where a substantial number of large datasets exist.

The aforementioned methods traditionally have only dealt with unimodal data, but have also become popular in multimodal emotion recognition, in which the detection of emotion in each modality is a critical component for the success of the entire multimodal system. One of the key challenges in multimodal emotion recognition is to model the interactions between each modality (i.e. inter-modality dynamics) (Marechal et al., 2019). While novel approaches (Zadeh et al., 2018) have been proposed to address this problem, a majority of the earlier work has been focused on the inter-modality dynamics within a single individual. In psychology, numerous studies (Parkinson, 2014; Smith, Alkozei, & Killgore, 2017) affirm that clearly another person's emotions do have an effect on our own actions, thoughts, and feelings. For instance, Paul Ekman (Ekman, Freisen, & Ancoli, 1980) highlights how one person's face may influence the emotional experience of another: "If B perceives A's facial expression of emotion, B's behavior toward A may change, and A's notice of this may influence or determine A's experience of emotion" (Ekman et al., 1980). In the field of multimodal emotion recognition on the other hand, little research has been done to explore the inter-modality dynamics between individuals (i.e. inter-personal). This research aims to further understand such inter-modality and inter-personal effects. Through an empirical study, we investigate the correlations between emotions extracted from the speech and physiological signals of a group of individuals, and the emotions from the facial expressions of another group.

3. Methodology

3.1 Data Collection

We collected physiological, visual, and audio data from both actors and audience during a theatre performance that took place in the Landesmuseum in Zurich, Switzerland on May 25th, 2019. Through the Happimeter app (Budner, Eirich, & Gloor, 2017) running on the smartwatch that the actors wore during the performance, we were able to gather the activation, pleasance, and stress levels of the actors. The Happimeter runs a trained machine learning model that is capable of predicting such emotions from the physiological signals that are collected from the sensors of the smartwatch. Through a video camera that was set-up inside the theatre, we captured the faces of the audience and the voices of the actors during the entire performance which lasted for about 40 minutes.

Considering the number of smartwatches available as well as the privacy issues imposed by the collection of sensor data, we opted to collect the physiological signals from the smaller group of individuals - the actors, which consisted of 8 individuals. Moreover, as a theatre etiquette, loud whispers and conversations in the audience are discouraged, hence, speech data was only collected from the actors. Facial expressions on the other hand, were recorded from the audience, which consisted of about 40 people. **Table 1** summarizes the data collected during the theatre performance.

Table 1. *Summary of the collected data from the theatre performance*

Modality	Emotions	Group
Facial expressions	Anger, Fear, Happiness, Sadness	Audience
Physiological signals	Activation, Pleasance, Stress	Actors
Speech signals	Anger, Fear, Happiness, Sadness	Actors

3.2 Model Implementation

3.2.1 Face Emotion Recognition

Our FER model, which has a prediction accuracy of 74.9%, has been trained on a combination of multiple datasets: CK+ (Lucey et al., 2010), JAFFE (Lyons, Kamachi, & Gyoba, 1998), BU-3DFE (Yin, Wei, Sun, Wang, & Rosato, 2006), and FacesDB (Mena-Chalco, Marcondes, & Velho, 2008). The cardinality of each emotion type in these datasets is summarized in **Table 2**. Our model uses a VGG16 (Simonyan & Zisserman, 2014) CNN architecture that was pre-trained on ImageNet. We freeze the layers except for the last 4 layers of this pre-trained model. We use SGD as an optimizer with a learning rate of 0.01 and a Softmax activation function in the dense output layer of the network. All of the detected faces from the camera were resized to 100 x 100 as input to the VGG16 model. Since VGG16 expects three input channels, we extend the images into three dimensions by using the same values for red, green, and blue (i.e. grayscale).

Table 2*. FER Training Set*

	Dataset				
Emotions	CK+	JAFFE	BU-3DFE	FacesDB	**Total**
Happy	69	31	77	36	213
Sad	28	31	88	36	183
Angry	45	30	94	35	204
Fearful	25	32	92	36	185
Total	167	124	351	143	**785**

Using the face_recognition python package which is based on the dlib machine learning library (King, 2009), we detect the faces from the captured images on the camera. We then label all the recognized faces with the emotions happy, angry, sad,

and fearful, using our trained FER model. The probabilities of the emotion classes are obtained from the Softmax layer of our FER network. Fig. 1 illustrates a snapshot of the video captured by the camera, showing the emotion-labeled faces as predicted by our FER model.

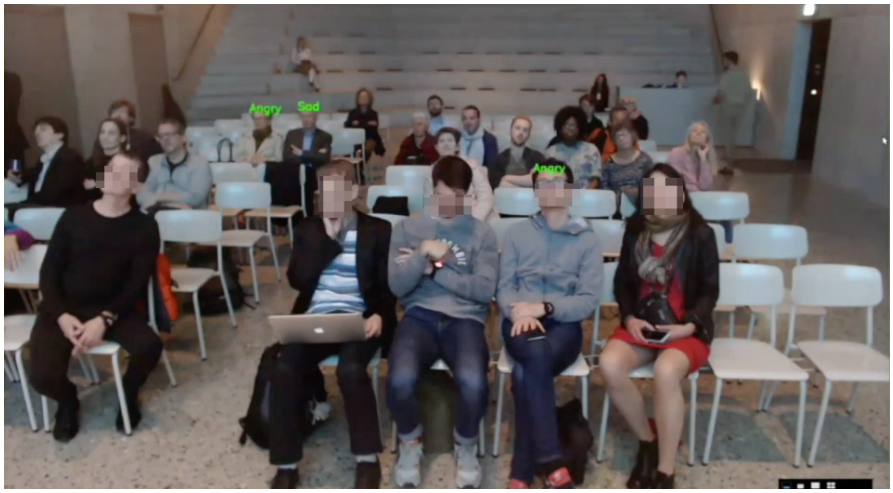


Fig. 1. *Emotion-labeled faces detected by our FER Model (face blurred for privacy reasons)*

3.2.2 Speech Emotion Recognition

Our SER model, which has a prediction accuracy of 71.01%, has been trained on a combination of multiple datasets containing 3-5 seconds of emotion-labeled audio files: RAVDESS (Livingstone & Russo, 2018), SAVEE (Jackson & ul haq, 2011), CREMA-D (Cao et al., 2014), IEMOCAP (Busso et al., 2008), TESS (Dupuis & Pichora-Fuller, 2011), and EMODB (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). The cardinality of each emotion type in these datasets is summarized in **Table 3**. Using python's Librosa (McFee et al., 2015) library, we extract the MFCCs (Mel-frequency cepstral coefficients) from each audio file with a sampling rate of 44100 Hz, a Fast Fourier Transform (FFT) window of 2048, and hop length of 512 samples. This implementation uses the Hann window function on the signal frames and performs a Short-Time Fourier-Transform (STFT) to calculate the frequency spectrum. We extract a total of 40 MFCCs, excluding the zeroth coefficient as it represents the average log-energy of the signal, which carries limited speech information (Nandi & Rao, 2015). We then feed this MFCC feature vector into our LSTM: a five-layer network with one input layer, 3 hidden layers, and one dense output layer with a Softmax activation function.

Table 3. *SER Training Set*

	Dataset						
Emotions	RAVDESS	SAVEE	CREMA-D	IEMOCAP	TESS	EMODB	**Total**
Happy	376	60	1271	595	400	72	2774
Sad	376	60	1271	1084	400	62	3253
Angry	376	60	1271	1103	399	128	3337
Fearful	376	60	1271	40	399	68	2214
Total	1504	240	5084	2822	1598	330	**11578**

Using the video captured by the camera, we extract the corresponding audio data by converting the mp4 into a wav file format. The entire audio stream was split, with a chunk length of 4 seconds, since our SER prediction model was trained on audio data with a similar average time duration. We then label each of the 4-second audio with the emotions happy, angry, sad, and fearful, using our trained SER model. The probabilities of the emotion classes are obtained from the Softmax layer of our SER network.

3.2.3 Physiological Emotion Recognition

We use the machine learning model that is deployed in the Happimeter (Budner et al., 2017) app to label the emotions from the physiological signals, i.e. a Physiological Emotion Recognition (PER). Signals were collected by the sensors of the smartwatch. The model processes physiological (e.g. movement, heart rate, etc.) and environmental (noise, weather, etc.) variables as inputs to a classifier. It uses Scikit-learn's (Pedregosa et al., 2011) Gradient Boosting algorithm with a learning rate of 0.1 and a maximum depth of 8 nodes in each tree. This machine learning model, which currently has a prediction accuracy of 79%, has been trained with the data that has been acquired from the users of the app from the past three years. Using this trained model, the data collected from the smartwatches that were worn by the actors, were labeled with values ranging from 0 to 2 to indicate the levels of activation, pleasance, and stress.

3.2.4 Correlation Analysis

We compare the predicted emotions from the voices and physiological signals of the actors to the emotions from the facial expressions of the people in the audience. We merge the predictions from our SER (actors) and FER (audience) models based on the closest timestamp and perform a rolling window calculation (i.e. simple moving average) using different time windows to filter out noise and expose the underlying properties of the curves. Subsequently, we perform a correlation analysis using Pearson's Correlation Coefficient (see Equation below), where n is the sample size, x_i and y_i are the individual sample points i, and $\bar{x}$ and $\bar{y}$ are the sample mean. The same process is followed to compare the emotions from the PER (actors) and the FER (audience) model. We also analyze the physiological signals (i.e. heartrate and movement) from the actors and examine their correlations with the emotions portrayed from the faces of the audience.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4. Results

4.1 PER vs FER

Fig. 2 shows the levels of activation, pleasance, and stress of the actors (as measured by the Happimeter app) and the four emotions of the audience (as measured by the FER model) throughout the entire theatre performance. As we can see, the pleasance of the actors went down as the play progressed, while their activation went up. The correlation values and the level of significance between these emotions are illustrated in **Fig. 3** (* <0.05, ** <0.01, *** <0.001).

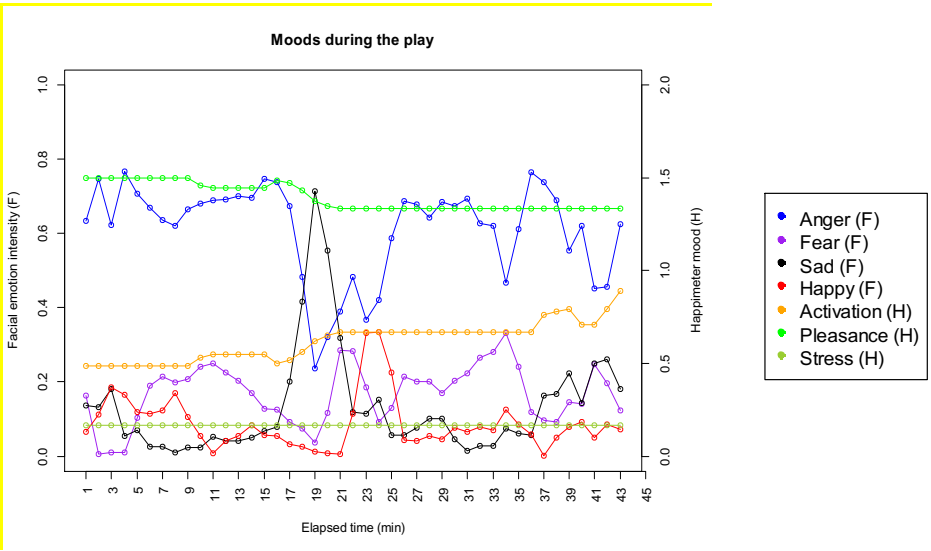


Fig. 2. *Emotions from the Happimeter and the FER model*

Fig. 3. *Correlations between the emotions from the Happimeter (actors, about 900 measurements) and the FER model (audience, about 600 measurements)*

We find that activation of the actors and anger of the audience is negatively correlated (r=-0.31*). This means that the more excited the actors are, the less angry the audience is. We do not really assume that the audience is "angry", rather their facial expressions showed something that our FER interpreted as "angry". As we only had these four emotions labeled this initial analysis, other emotions such as "surprise" or "insight" might be subsumed into the "angry" emotion, as the FER system might assign these emotions also the "angry" label. Similarly, we find that the higher the pleasance of the actors is, the less "fearful" the audience is (r=-0.32*). Somewhat counterintuitively we also find that the higher the pleasance of the actors is, the more angry the audience (r=0.39**) is. This combination of correlations indeed suggests that the "surprise" facial expression might be similar to the "anger" facial expression and has been recognized as such by the FER.

4.2 *Sensor data vs FER*

In order to investigate the possible correlations between raw sensor data (as captured by the smartwatch) and the FER model, we collected and analyzed the data from the smartwatches worn by the actors. ***Fig. 4.*** *shows the average levels of movement (computed as the sum of the absolute values of accelerometers value along x, y and z-axis), heartrate (beats per minute, BPM), and noise level (as measured by microphone). The correlation*

values and the level of significance between these emotions are illustrated in

	MVMT_avg	BPM_avg	mic_avg	angry	fear	sad	happy
MVMT_avg	1	*** -0.52	0.2	** -0.42	* 0.36	0.13	0.18
BPM_avg		1	-0.06	0.13	-0.26	0.06	-0.05
mic_avg			1	-0.28	-0.03	* 0.3	-0.01
angry				1	* -0.34	*** -0.71	* -0.32
fear					1	-0.2	-0.21
sad						1	-0.07
happy							1

Fig. 5.

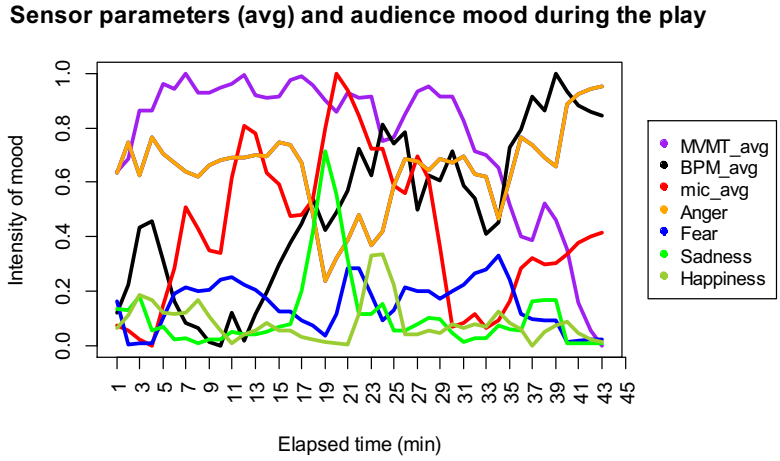
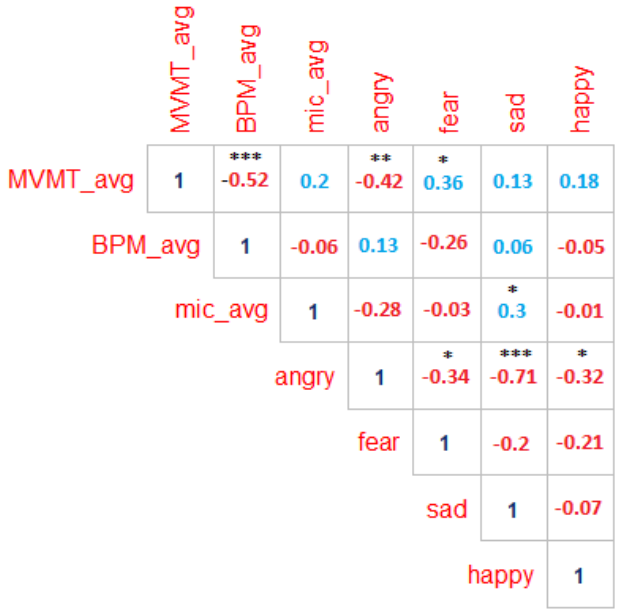


Fig. 4. *Sensor data (average) comparison with the FER model*



	MVMT_avg	BPM_avg	mic_avg	angry	fear	sad	happy
MVMT_avg	1	*** -0.52	0.2	** -0.42	* 0.36	0.13	0.18
BPM_avg		1	-0.06	0.13	-0.26	0.06	-0.05
mic_avg			1	-0.28	-0.03	* 0.3	-0.01
angry				1	* -0.34	*** -0.71	* -0.32
fear					1	-0.2	-0.21
sad						1	-0.07
happy							1

As

Fig. 5 shows, the facial expression recognized as "angry" is negatively correlated to the average movement, i.e. the less the actors move, the more "angry" the audience gets.

	MVMT_avg	BPM_avg	mic_avg	angry	fear	sad	happy
MVMT_avg	1	*** -0.52	0.2	** -0.42	* 0.36	0.13	0.18
BPM_avg		1	-0.06	0.13	-0.26	0.06	-0.05
mic_avg			1	-0.28	-0.03	* 0.3	-0.01
angry				1	* -0.34	*** -0.71	* -0.32
fear					1	-0.2	-0.21
sad						1	-0.07
happy							1

Fig. 5. *Correlations between sensor data (average, about 2150 measurements) and the FER model*

Similarly, we find, that the higher the standard deviation in movement of the actors, the "angrier" expressions (r=0.32*) and the less "fear" expressions (r=-0.34*) are recognized by the FER. This means that differences in movement among the actors trigger emotional reactions by the audience.

4.3 FER vs SER

Fig. 6 shows the plots of the probabilities of the emotion "anger" as measured by our FER and SER models using a rolling time window of 30 seconds, one minute, and five minutes. As foreseen, a smoother curve is achieved with a longer time window. In Fig. 7 the plots of the probabilities of all four emotions between the actors (as predicted by the SER) and the audience (as predicted by the FER) with a rolling window of one minute are displayed. The corresponding correlation matrix showing the correlation values and the level of significance is displayed in **Fig. 8**. Only significant correlations between the emotions of the audience and actors (i.e. FER vs SER predictions) are highlighted.

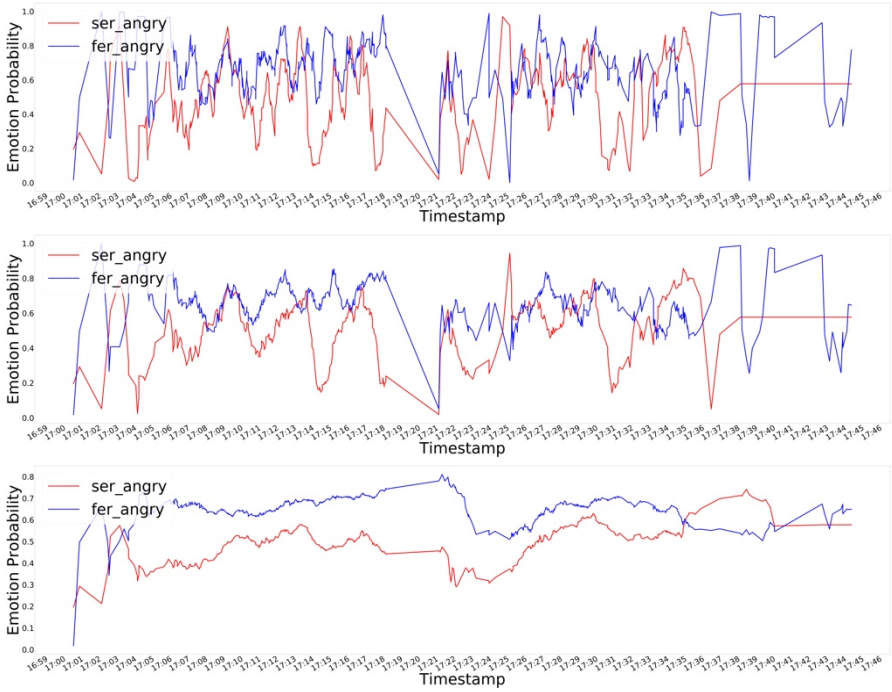


Fig. 6. *Plot of the probabilities of the emotion anger from the FER and SER models using different time windows*

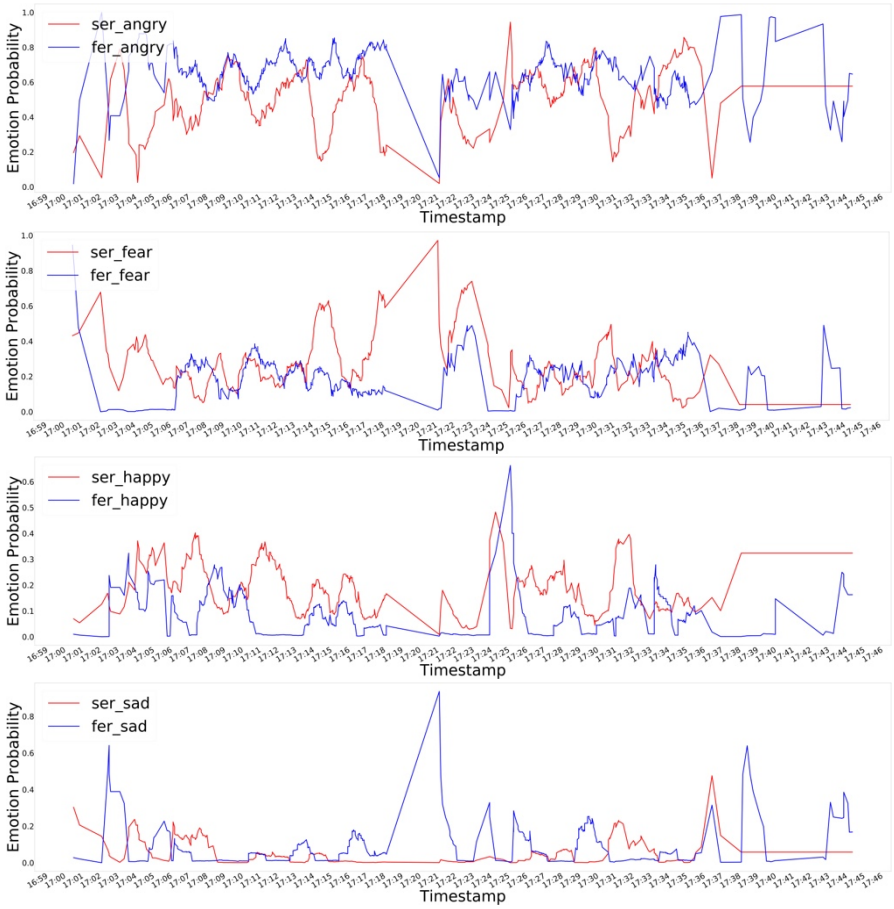


Fig. 7. *Plot of the probabilities of the four emotions from the FER and SER models using a rolling window of one minute*

As the correlation matrix in Fig. 8 shows, "fear" in the faces of the audience is positively correlated with "anger" in the voice of the actors. "Anger" in the faces of the audience is positively correlated with "happiness" in the voice of the actor, which again suggested that "surprise" of the audience is also subsumed in this emotion.

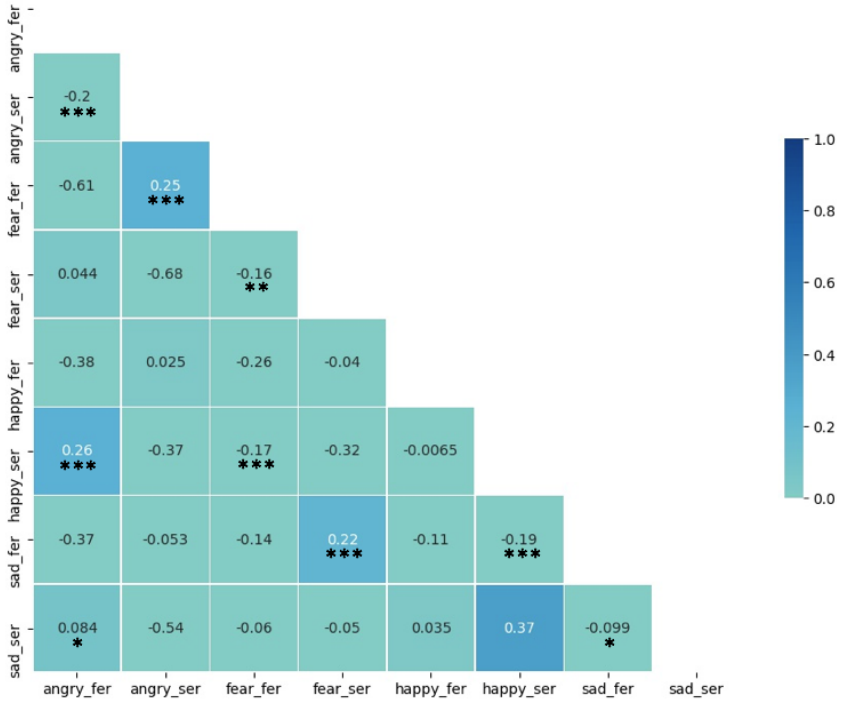


Fig. 8. *Correlation matrix between the emotions from the FER (N= 592) and SER (N= 684) models*

5 Discussion

5.1 Emotions from Faces of the Audience vs Voices of Actors

By taking into consideration a balance between filtering random noise or variations and preserving the original data, we chose a time window of one minute to smoothen the time series predictions as can be observed from the plots in **Fig. 6**. Using this chosen time window, we see some obvious correlations between the emotions from the audience and the actors (see **Fig. 8**). A graphical summary of the correlations are shown in Fig. 9, which is based on the correlation matrix in Fig. 8.

For the emotion "anger", there is a statistically significant negative correlation between the audience and the actors. Interestingly, there is a statistically significant positive correlation between the "happiness" from the actors and "anger" from the audience. This implies that when there is "anger" from the actors, the audience feels less of the same emotion and similarly, when there is "happiness" from the actors, there is a higher intensity of "anger" from the audience.

The “anger” expressed by the voices of the actors is positively correlated with “fear” from the audience, which appear to be logical and can possibly infer that the actors can effectively elicit “fear” from the audience by demonstrating “anger” in their voices. Consistent with such behavior, there is also a statistically significant negative correlation between the “happiness” from the actors and “fear” from the audience, implying that the audience feels less “fear” when the actors exhibit “happiness”.

A statistically significant positive correlation is also present between “fear” from the actors and “sadness” from the audience. This may suggest that members of the audience are sympathetic, and they empathize with the “fear” from the actors by feeling “sad”. Consistent with such observation, there is also a statistically significant negative correlation between “happiness” from the actors and “sadness” from the audience, which suggests that the actors effectively managed to make the audience feel less “sad” by showing “happiness” through their voices.

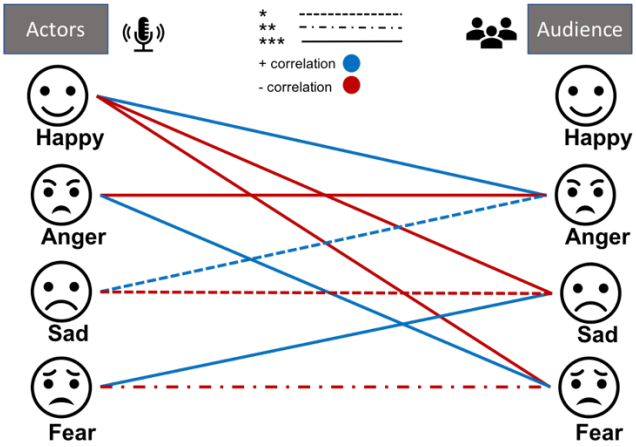


Fig. 9. *Significant correlations between the actors and the audience based on the FER and SER correlation matrix*

5.2 Emotions from Happimeter vs Faces of the Audience

Based on the same considerations as discussed in 5.1, we chose a time window of three minutes to smoothen the time series predictions of Happimeter and FER as shown in the plots in Figure 6. In this plot, the actors’ emotions “pleasance” and “activation” are compared with the audience’s “angry”, “fear”, “sad” and “happy” facial expressions.

We find that the variable “angry” is negatively correlated to the “activation” of Happimeter and positively correlated to “pleasance”. This seems to suggest that the angry emotion is covering another emotion (maybe “surprise”) as it leads the

audience to be more agitated. As expected, the audience variable "fear" is negatively correlated to the actors' "pleasance".

5.3 Actors' Sensor Data vs Faces of the Audience

We find that an increase of the average "movement" of the actors leads to a decrease of "angry" emotions among the audience, in accordance with the discussion in 5.2, but also to an increase of the "fear" emotion. Moreover, an increase of the average sound level measured with the microphone is positively correlated to the "sadness" of the audience. We assume that this is directly related to the theater piece which was played in this analysis, where tragic experiences of the protagonist are presented.

We also observed that an increase in the variance of movements leads to an increase in the anger of the audience, while decreasing their fear. This might be related to one actress walking among the audience, triggering some anger and fear of spectators of being called out.

A graphical summary of the correlations discussed in 5.2 and 5.3 are shown in Fig. 10, which is based on the correlation matrices in Fig. 8 and Fig. 5.

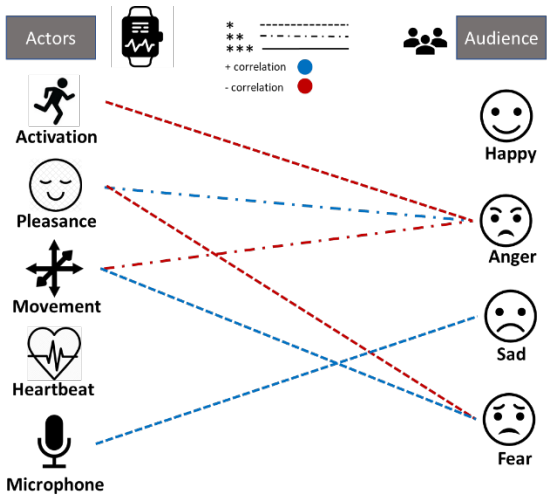


Fig. 10. *Significant correlations between the actors and the audience based on the PER vs FER and Sensor data vs FER correlation matrices.*

6 Conclusions and Future Work

One of the main restrictions of the analysis described in this paper is that the FER we used only is capable of recognizing the four emotions happy, sad, fear, and anger, potentially leading to overrecognition of fear and anger. In the revised version of the FER which has been developed in the meantime, we have included the two additional emotions of the Ekman model surprise and disgust, which in more recent work have shown increased recognition accuracy and emotion coverage.

Nevertheless, we are convinced that the system described in this paper has illustrated the potential of our approach of automatically measuring audience and artist emotions at public events. We are currently extending our system for using it at other artistic events such as concerts and other public events. In particular this includes giving immediate feedback to participants about their emotions, and combining sound input from other sources such as smartphones with the Happimeter and the video input from the Webcam. Our ultimate goal will be to identify the emotions that will lead to optimal experiences for both performers and the audience. Mirroring back this behavior (Gloor et al. 2017) to performers will allow them to better understand the impact their own emotions have on their audience, and thus to improve their artistic performance and skills.

Acknowledgements

We thank Samuel Schwarz and Garrick Lauterbach from Digitalbuehne.ch for providing the venue and supporting us during the performance. We are also grateful to Jannik Roessler for his invaluable support during data collection.

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional/ and or national research committee with the Helsinki declaration and its later amendments or comparable with ethical standards. Consent by the organizers and participants in the Zurich experiment to be recorded has been given.

References

Albanie, S., Nagrani, A., Vedaldi, A., & Zisserman, A. (2018). Emotion recognition in speech using cross-modal transfer in the wild. *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, 292–301. https://doi.org/10.1145/3240508.3240578

Ali, M., Mosa, A. H., Machot, F. Al, & Kyamakya, K. (2018). A Review of Emotion Recognition Using Physiological Signals. *Annals of Telecommunications -*

Annales Des Télécommunications, *109*(3–4), 303–318. https://doi.org/10.1007/978-3-319-58996-1

Ambady, N., & Weisbuch, M. (2010). Nonverbal behavior. In *Handbook of social psychology, Vol. 1, 5th ed.* (pp. 464–497). Hoboken, NJ, US: John Wiley & Sons Inc.

Budner, P., Eirich, J., & Gloor, P. A. (2017). "Making you happy makes me happy" - Measuring Individual Mood with Smartwatches, (Aristotle 2004), 1–14. Retrieved from http://arxiv.org/abs/1711.06134

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). *A database of German emotional speech. 9th European Conference on Speech Communication and Technology* (Vol. 5).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., … Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335. https://doi.org/10.1007/s10579-008-9076-6

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Chen, M., Zhang, Y., Qiu, M., Guizani, N., & Hao, Y. (2018). SPHA: Smart Personal Health Advisor Based on Deep Analytics. *IEEE Communications Magazine*, *56*(3), 164–169. https://doi.org/10.1109/MCOM.2018.1700274

Dupuis, K., & Pichora-Fuller, M. (2011). Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Canadian Acoustics - Acoustique Canadienne*, *39*, 182–183.

Egger, M., Ley, M., & Hanke, S. (2019). Emotion Recognition from Physiological Signal Analysis: A Review. *Electronic Notes in Theoretical Computer Science*, *343*, 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

Ekman, P., Freisen, W. V, & Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology*, *39*(6), 1125–1134. https://doi.org/10.1037/h0077722

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*. US: American Psychological Association. https://doi.org/10.1037/h0030377

Gloor, P., Colladon, A. F., Giacomelli, G., Saran, T., & Grippa, F. (2017). The impact of virtual mirroring on customer satisfaction. *Journal of Business Research*, *75*, 67-76.

Hong, W., Zheng, C., Wu, L., & Pu, X. (2019). Analyzing the relationship between consumer satisfaction and fresh e-commerce logistics service using text mining techniques. *Sustainability (Switzerland)*, *11*(13), 1–16. https://doi.org/10.3390/su11133570

Jackson, P., & ul haq, S. (2011, April 1). Surrey Audio-Visual Expressed Emotion (SAVEE) database.

Karpathy, A., & Li, F.-F. (2014). Deep Visual-Semantic Alignments for Generating Image Descriptions. *CoRR*, *abs/1412.2*. Retrieved from

http://arxiv.org/abs/1412.2306

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors (Switzerland)*, *18*(2). https://doi.org/10.3390/s18020401

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (pp. 1097–1105). USA: Curran Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2999134.2999257

Lee, J., & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015-Janua*, 1537–1540.

Li, Y., Zhu, J.-Y., Tedrake, R., & Torralba, A. (2019). Connecting Touch and Vision via Cross-Modal Prediction, (d). Retrieved from http://arxiv.org/abs/1906.06322

Livingstone, S. R., & Russo, F. A. (2018, April). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). https://doi.org/10.5281/zenodo.1188976

Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). *The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression*. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*. https://doi.org/10.1109/CVPRW.2010.5543262

Lyons, M., Kamachi, M., & Gyoba, J. (1998, April). The Japanese Female Facial Expression (JAFFE) Database. Zenodo. https://doi.org/10.5281/zenodo.3451524

Marechal, C., Mikołajewski, D., Tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C., & Węgrzyn-Wolska, K. (2019). Survey on AI-based multimodal methods for emotion detection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11400*, 307–324. https://doi.org/10.1007/978-3-030-16272-6_11

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference*, (Scipy), 18–24. https://doi.org/10.25080/majora-7b98e3ed-003

Mena-Chalco, J., Marcondes, R., & Velho, L. (2008). *Banco de Dados de Faces 3D: IMPA-FACE3D*.

Nandi, D., & Rao, K. (2015). *Language Identification Using Excitation Source Features*. https://doi.org/10.1007/978-3-319-17725-0

Owens, A., Isola, P., McDermott, J. H., Torralba, A., Adelson, E. H., & Freeman, W. T. (2015). Visually Indicated Sounds. *CoRR*, *abs/1512.0*. Retrieved from http://arxiv.org/abs/1512.08512

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition, (Section 3), 41.1-41.12. https://doi.org/10.5244/c.29.41

Parkinson, B. (2014). How Emotions Affect Other People. *Emotion Researcher*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, *12*, 2825–2830. Retrieved from http://dl.acm.org/citation.cfm?id=1953048.2078195

Posner, J., Russell, J., & Peterson, B. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*, 715–734. https://doi.org/10.1017/S0954579405050340

Qiu, J. L., Liu, W., & Lu, B. L. (2018). Multi-view emotion recognition using deep canonical correlation analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11305 LNCS*, 221–231. https://doi.org/10.1007/978-3-030-04221-9_20

Rule, N., & Ambady, N. (2010). First Impressions of the Face: Predicting Success. *Social and Personality Psychology Compass*, *4*(8), 506–516. https://doi.org/10.1111/j.1751-9004.2010.00282.x

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Smith, R., Alkozei, A., & Killgore, W. (2017). How Do Emotions Work? *Frontiers for Young Minds*, *5*. https://doi.org/10.3389/frym.2017.00069

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*(1), 93–120. https://doi.org/10.1007/s10772-018-9491-z

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., … Rabinovich, A. (2014). Going Deeper with Convolutions. *CoRR*, *abs/1409.4*. Retrieved from http://arxiv.org/abs/1409.4842

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2016-May*, 5200–5204. https://doi.org/10.1109/ICASSP.2016.7472669

Ullah, M. A., Islam, M. M., Azman, N. B., & Zaki, Z. M. (2017). An overview of Multimodal Sentiment Analysis research: Opportunities and Difficulties. *2017 IEEE International Conference on Imaging, Vision and Pattern Recognition, IcIVPR 2017*. https://doi.org/10.1109/ICIVPR.2017.7890858

Vasey, E., Ko, S., & Jeon, M. (2018). In-Vehicle Affect Detection System: Identification of Emotional Arousal by Monitoring the Driver and Driving Style. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 243–247). New York, NY, USA: ACM. https://doi.org/10.1145/3239092.3267417

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys*, *50*(2), 1–33. https://doi.org/10.1145/3057270

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3D Facial Expression Database For Facial Behavior Research. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition* (pp. 211–216). Washington, DC, USA: IEEE Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=1126250.1126340

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2018). Tensor Fusion Network for Multimodal Sentiment Analysis, 1103–1114. https://doi.org/10.18653/v1/d17-1115