INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

OpenMM: An Open-source Multimodal Feature Extraction Tool

Michelle Renee Morales[1], *Stefan Scherer*[2], *Rivka Levitan*[3]

[1]Department of Linguistics, CUNY Graduate Center, USA
[2]Institute for Creative Technologies, University of Southern California, USA
[3]Department of Computer Science, Brooklyn College (CUNY), USA

mmorales@gradcenter.cuny.edu, scherer@ict.usc.edu, levitan@sci.brooklyn.cuny.edu

Abstract

The primary use of speech is in face-to-face interactions and situational context and human behavior therefore intrinsically shape and affect communication. In order to usefully model situational awareness, machines must have access to the same streams of information humans have access to. In other words, we need to provide machines with features that represent each communicative modality: face and gesture, voice and speech, and language. This paper presents OpenMM: an open-source multimodal feature extraction tool. We build upon existing open-source repositories to present the first publicly available tool for multimodal feature extraction. The tool provides a pipeline for researchers to easily extract visual and acoustic features. In addition, the tool also performs automatic speech recognition (ASR) and then uses the transcripts to extract linguistic features. We evaluate the OpenMM's multimodal feature set on deception, depression and sentiment classification tasks and show its performance is very promising. This tool provides researchers with a simple way of extracting multimodal features and consequently a richer and more robust feature representation for machine learning tasks.

Index Terms: feature engineering, multimodal, classification

1. Introduction

When focusing on classification tasks, researchers across various fields—speech processing, Natural Language Processing, and human-computer interaction—often work on the same task from different perspectives, usually with different data sources and feature representations. However, to truly get a comprehensive picture of a conversation, it is necessary to consider all modalities. In many situations, a multimodal system can provide the most robust source of information for a classification task and research has found that on average multimodal systems offer an 8% improvement over unimodal systems [1]. However, building a multimodal system is extremely time intensive because it requires feature engineering across multiple modalities. In some cases, it is also not feasible given the dataset type. The goal of OpenMM is to provide researchers with a simple tool to extract multimodal features. OpenMM is built upon various existing open-source tools as well as our own code for linguistic analysis. The tool only requires a video as input and performs all the processing and necessary conversions to generate audio files and transcriptions, as shown in Figure 1.

Given a video input, OpenMM will extract visual features using the open-source tool OpenFace [2]. Then OpenMM converts the video to audio using the tool ffmpeg [3], outputting an audio wav file. Using the wav file, OpenMM extracts acoustic features using the open-source repository Covarep [4]. Using the wav file, OpenMM then makes a call, depending on the language, to either IBM Watson's or Google's speech-to-text service, outputting a transcript. Using the transcript OpenMM then extracts linguistic features, which include a bag-of-words representation and syntactic features. The syntactic features are generated using a dependency parse tree representation, which is generated using Google's state-of-the-art parser. In the end, OpenMM outputs the following: wav file, transcript file, comma-separated values file (CSV) of visual features, CSV of acoustic features, CSV of linguistic features, and CSV of multimodal features. OpenMM currently supports English, German, and Spanish and is available for download[1].

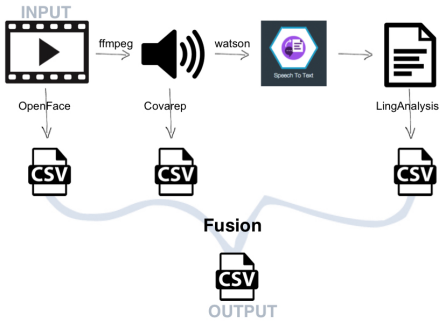

Figure 1: *OpenMM Pipeline*

We evaluate the OpenMM multimodal feature set on three different classification tasks. The three classification tasks we consider are deception (deceptive vs. truthful), depression (depressed vs. not depressed), and sentiment detection (negative vs. positive), which respectively involves datasets in English, German, and Spanish. In many experiments, we find OpenMM features match or outperform previous systems [5, 6]. Using OpenMM we are able to classify deception with 76.86% accuracy, sentiment with 62.50% accuracy, and depression with 76.79%. We hope this tool will provide researchers with a simple and inexpensive way of extracting multimodal features.

2. Related Work

In this section we provide brief overviews of related work on the three relevant classification tasks.

2.1. Deception Detection

Pérez-Rosas et al. [6] presented the first multimodal system to detect deception in real-life trial data using text and gesture modalities. They built classifiers relying on individual and combined sets of nonverbal and verbal features, reporting accuracies in the range of 60-75%. Their dataset was manually transcribed and their verbal features included unigrams and bigrams derived from the bag-of-words representation of their video transcripts.

[1]https://github.com/michellemorales/OpenMM

http://dx.doi.org/10.21437/Interspeech.2017-1382

Their nonverbal features included facial displays and hand gestures, which were also manually annotated by hand. Pérez-Rosas et al.'s findings show the promise in multimodal features while also highlighting the time intensive nature of multimodal design, which often includes a good deal of manual annotation.

2.2. Depression Detection

Researchers have also investigated the use of multimodal features for depression detection. Recent research has shown the promise in using acoustic [7, 8, 9, 10, 11] and visual features [12, 13, 11] for depression detection. Some researchers have built multimodal systems, specifically Scherer et al. [14], investigated visual signals and voice quality, finding that they were able to distinguish interviewees with depression from those without depression with an accuracy of 75%. In addition to audiovisual features, text-based features, including syntactic and semantic features have been investigated for depression detection. Rude et al. [15] examined linguistic patterns of student narratives, finding that depressed students used significantly more first person singular words and negatively valenced words than did never-depressed students. In addition to considering word usage, researchers have explored syntactic characteristics of depressed language [16, 17]. Zinken et al. [16] investigated whether an analysis of a depressed patients' syntax could help predict improvement of symptoms and found that certain syntactic structures were correlated with patients' potential to complete a self-help treatment.

2.3. Sentiment Detection

In other work, Pérez-Rosas et al. [5] presented a method for multimodal sentiment classification, which could identify the sentiment of video reviews. In order to identify sentiment, they explored visual, acoustic, and text features. Acoustic features were extracted using the open-source software OpenEAR [18], which extracts prosody, energy, voice probabilities, spectrum, and cepstral features. Facial features were extracted using the Computer Expression Recognition Toolbox [19], including smile and head pose estimates, facial action units, and eight basic emotions. Lastly, video clips were manually processed to transcribe the verbal statements and to extract the start and end time of each utterance. In addition, linguistic features were extracted from the transcripts using a bag-of-words representation. Their work showed that multimodal sentiment analysis can be effectively performed. They also report that the joint use of multimodal features (visual, acoustic, and linguistic) can lead to error rate reductions of up to 10.5% as compared to the best performing single modality.

3. Datasets

In this work, we use three publicly-available datasets: the Audio-Visual Emotion Recognition Challenge (AVEC) dataset[2], the Multimodal Opinion Utterances dataset (MOUD), and the Real-life Trial (RLT) dataset[3].

3.1. Audio-Visual Emotion Recognition Challenge Dataset

For the depression classification task, we use the 2014 AVEC corpus [20]. In total, the corpus includes 300 videos in German. Since we are concerned with spontaneous language, we only use half of the corpus from the spontaneous 'freeform' speech task. In total, this subset of the corpus is composed of 150 videos. The videos include recordings of participants responding to one of a number of questions. Each recording is labeled for severity of depression. Depression severity is determined using the Beck Depression Inventory-II (BDI-II) [21]. BDI-II scores range from 0 to 63. We group the data into 2 binary classes not depressed (0-13) and depressed ($\geq$14).

3.2. Multimodal Opinion Utterances Dataset

The dataset we use for sentiment classification is the MOUD dataset [5]. The MOUD dataset includes videos of product opinions expressed in Spanish. The videos were collected from the social media web site YouTube, using several search keywords that were likely to lead to product reviews or recommendations. Videos selected met the following guidelines: the speaker was directly in front of the camera, his/her face was clearly visible, with minimum amount of face occlusion, and no background noise. In total the dataset is comprised of 80 videos randomly selected from the videos retrieved from YouTube that met the guidelines. All video clips were manually processed to transcribe the verbal statements and to extract the start and end time of each utterance. Each utterance was then labeled for sentiment by two annotators.

3.3. Real-life Trial Dataset

The RLT dataset [6] includes videos of real deception during court trials in English. Videos were collected from public multimedia resources where trial hearing recordings were available and where truthful or deceptive behavior could be fairly observed and verified. Videos selected met the following guidelines: defendant or witness in the video should be clearly identified, his/her face should be visible during most of the clip duration, visual quality should be clear enough to identify facial expressions, and clear audio quality. Trial outcomes, such as guilty verdict, non-guilty verdict and exoneration, are used to help correctly label video clips with deceptive or truthful.

4. OpenMM Feature Extraction

OpenMM aims to extract features from as many channels or modalities as possible, including nonverbal behavior, voice and speech characteristics, as well as linguistic characteristics.

4.1. Automatic Speech Recognition

Given advancements in ASR, language can now be a common component in classification systems. For this reason, it is important to investigate how successful a feature set can be when it is fully automated. Manual transcription ensures the most accurate transcription possible, however it is expensive in time and resources. Therefore, it is important to investigate how ASR transcript derived features compare to manual transcription derived features. OpenMM includes ASR to automate the transcription process. For English and Spanish, we use Watson's Speech-to-Text API [4]. For German, we use Google's API [5]

[2] https://avec2013-db.sspnet.eu/

[3] The MOUD and the RLT datasets can be found here: http://web.eecs.umich.edu/ mihalcea/downloads.html

[4] www.ibm.com/watson/developercloud/doc/speech-to-text/

[5] https://cloud.google.com/speech/

Table 1: *Dependency distance measure.*

Dependency Relation	Distance
NSUBJ(saw-2, I-1)	1
DOBJ(man-4, saw-2)	2
DET(man-4, the-3)	1
PREP(with-5, man-4)	1
POBJ(glasses-6, with-5)	1
Sum	**6**

4.2. Verbal Features

4.2.1. Bag-of-words

Each ASR transcript represents a string of words, with no punctuation or capitalization included. We take each transcript and generate a bag-of-words representation of each sentence to derive unigram counts, which are then used as linguistic features. We first build a vocabulary consisting of all the words occurring in the transcriptions. We then remove words that have a frequency below 10, this threshold is based off previous work which found this threshold useful for deception and sentiment detection [5, 6]. The remaining words represent the unigram features. So for each sentence, we generate a feature vector that represents the frequency of the unigrams inside that utterance.

4.2.2. Syntax Features

In order to generate syntactic features, we first tag and parse all sentences, using Google's state-of-the-art pre-trained English parser: Parsey McParseface [22]. We also use Google's Spanish and German universal parsers. For each sentence S, the parser outputs universal part-of-speech (POS) tags. Grammatical roles are also labeled, which show how words in the sentence relate to one another. For example, the sentence "*I saw the man with glasses*" when parsed would output the dependency relationships listed in Table 1.

Using the parser's output, syntactic features are generated, including: depth of tree, number of root dependents, number of unique universal POS tags, frequency of each POS tag, average word length, and a computed dependency distance measure. The depth of the tree represents the number of levels in the tree, which gives a measure of how complex of a construction the sentence is. The number of root dependents represents the total number of children the root has, providing another way to represent sentence complexity. The number of unique POS tags captures how many unique POS tags were used, which measures syntactic variety. The average word length represents a simple way of capturing how advanced the vocabulary is. Lastly, the dependency distance measure is based on related work [23]. Given each parse tree, each dependency relation receives a distance score calculated as the absolute difference between the serial positions of the words that participate in the relation, i.e. difference between indices in the sentence. The dependency distance measure is then the sum of all the dependency distances in the sentence, as shown in Table 1.

4.3. Nonverbal Features

4.3.1. Facial Features

OpenFace [24] is used to extract 408 visual features. OpenFace is an open-source facial behavior analysis toolkit, which has achieved state-of-the-art results in facial landmark detection, head pose estimation, facial action unit recognition, and eye gaze estimation. OpenFace includes features that capture basic information about the video, such as frame number, timestamp, and confidence values. Features also include information about an individual's gaze as well as the location of their head and face, which are represented in the gaze, pose, and landmark features. In addition, OpenFace includes features from the Facial Action Coding System (FACS) [25]. FACS is a system used to taxonomize human facial movements by their appearance on the face. It is a commonly used tool and has become standard to systematically categorize physical expressions, which has proven very useful for psychologists and animators. FACS is composed of facial AUs (Action Units), which represent the fundamental actions of individual muscles or groups of muscles.

Table 2: *Deception accuracy reported for leave-one-out cross-validation using DT and RF algorithms.*

Feature Set	DT	RF
P2015 - Bag-of-words	60.33	56.19
P2015 - Facial	70.24	76.03
OpenMM - Bag-of-words	66.94	59.50
OpenMM - Syntax	57.02	62.81
OpenMM - Acoustic	75.21	76.86
OpenMM - Visual	71.07	73.55
P2015 - Verbal	60.33	50.41
P2015 - Nonverbal	68.59	73.55
OpenMM - Verbal	61.16	59.50
OpenMM - Nonverbal	74.38	75.21
P2015 - All Features	75.20	50.41
OpenMM - All	73.55	76.03

4.3.2. Acoustic Features

In order to extract features from the voice, we use Covarep (A Cooperative Voice Analysis Repository for Speech Technologies) [4]. Covarep is an open-source toolkit of advanced speech processing algorithms. Using Covarep we extract 71 audio features, including prosodic, source, and spectral features.

4.4. Fusion

For each unimodal feature set, OpenMM outputs a CSV of features. For the visual and acoustic features, the features are computed at the frame-level. For the text-based features, features are computed at the sentence-level. In order to fuse the modalities, we need one feature vector per modality. So, we apply statistical functionals to each unimodal feature set. We apply the following statistical functionals: maximum, minimum, mean, median, standard deviation, variation, kurtosis, skewness, 25% percentile, 50% percentile, and 75% percentile. Using the feature vector derived through the statistical functionals, we then fuse the modalities by concatenating each of the video-level feature vectors. In addition to the multimodal (verbal + nonverbal) feature set, we also fuse the verbal (bag-of-words + syntax) and nonverbal (acoustic + visual) modalities.

4.5. Experiments and Results

We conduct three series of experiments. For each series, we build and evaluate classification models using OpenMM's feature sets.

Table 3: *Sentiment classification accuracy reported for ten-fold cross-validation using SMO.*

Feature Set	SMO
P2013 - Verbal	73.33
P2013 - Acoustic	53.33
P2013 - Visual	50.66
OpenMM - Bag-of-words	48.96
OpenMM - Syntax	60.42
OpenMM - Acoustic	61.46
OpenMM - Visual	62.50
P2013 - Nonverbal	61.33
OpenMM - Verbal	52.08
OpenMM - Nonverbal	59.38
P2013 - All	74.66
OpenMM - All	57.29

4.6. Deception Detection

In order to compare directly to Pérez-Rosas et al.'s (2015) previous work on deception detection [6], we use the same experimental configuration. Therefore, we evaluate using two classification algorithms, Decision Trees (DT) and Random Forest (RF), using the Weka toolkit with default parameters [26]. We run several comparative experiments using leave-one-out cross-validation. In Table 2, we report our results in conjunction with Pérez-Rosas et al.'s results, which we refer to as P2015. Given the distribution between deceptive and truthful clips, the baseline on this dataset is 50.4%. We find that the deception prediction accuracy for OpenMM's multimodal feature set is 76.03% which matches P2015's best performing system. These results are extremely promising as they confirm that OpenMM's fully automated system can match the performance of a manually handcrafted feature set. In addition, across modalities OpenMM's performance matches or outperforms P2015's models. This is especially interesting in regards to verbal features; OpenMM's bag-of-words, syntax, and verbal feature sets outperform P2015's verbal feature sets. These findings confirm that ASR transcript derived features can compete with manual transcription derived features. Lastly, the OpenMM acoustic feature set achieves the best results, classifying deception with 76.86% accuracy.

4.7. Sentiment Detection

Similar to the the deception detection experiments, we also compare OpenMM's sentiment detection results directly with previous work. Results for OpenMM's models can be compared directly to Pérez-Rosas et al.'s (2013) [5] systems' results, which we refer to as P2013 in Table 3. As before, we evaluate each unimodal feature set as well as multimodal feature sets. Following P2013, we use an SVM classifier in ten-fold cross-validation experiments. Given the distribution between positive and negative clips, the baseline on this dataset is 55.93%. For sentiment detection, we find that OpenMM's unimodal acoustic and visual features outperform P2013's feature sets. However, we also find that OpenMM's verbal feature sets are unable to match the performance of P2013. We think this can be attributed to the ASR model. Specifically for Watson's speech-to-text service, IBM announced that their English conversational speech recognition system achieves an 8% word error rate. However, they also mention having little data for building the Spanish model, leading to far higher error rates [27]. We believe this difference in ASR model performance led to poorer performing verbal feature sets. Although, we find ASR to be an extremely valuable tool for feature engineering for the English task, these results for Spanish highlight the limitations of ASR.

Table 4: *Depression classification accuracy reported for leave-one-out cross-validation using SMO.*

Feature Set	SMO
OpenMM - Bag-of-words	44.64
OpenMM - Syntax	44.64
OpenMM - Acoustic	76.79
OpenMM - Visual	62.50
OpenMM - Verbal	46.43
OpenMM - Nonverbal	62.50
OpenMM - All	62.50

4.8. Depression Detection

Lastly, we evaluate OpenMM's feature sets on the depression detection task. Results are given in Table 4. Given the distribution between depressed and not depressed clips, the baseline on this dataset is 55.36%. Since we only use half of the AVEC corpus and conduct a classification experiment, instead of the more common regression, it is difficult to provide a direct system comparison for depression detection. However, given the difficult nature of the task, we believe OpenMM's results show promise. The visual, acoustic, and multimodal features perform better than the baseline. As shown in Table 4, the OpenMM nonverbal, acoustic, and visual feature sets achieve the best results. The acoustic feature set represents the highest performing system, reporting an accuracy of 76.79%. These results confirm previous findings that acoustic and visual features are extremely useful for depression detection [28]. Similar to what we found for sentiment detection, the verbal feature sets represent the lowest performing systems, which is again likely an artifact of the German ASR model.

5. Conclusions

In this paper, we present OpenMM, the first open-source multimodal feature extraction tool. We evaluate OpenMM on three datasets spanning three different languages. We find that OpenMM's unimodal and multimodal feature sets perform well across different classification tasks. Our best performing models are able to classify deception with 76.86% accuracy, sentiment with 62.50% accuracy, and depression with 76.79% accuracy. Our findings show that multimodal features derived from a fully automated system can match the performance of a manually handcrafted feature set. In addition, we find that features derived from ASR transcriptions can compete with features derived from manual transcriptions. We hope OpenMM will provide researchers with a simple and inexpensive way of extracting multimodal features, which encompass various communicative modalities: face and gesture, voice and speech, and language. Lastly, we hope OpenMM can lead to richer and more robust feature representations for machine learning tasks.

6. References

[1] S. D'Mello and J. Kory, "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 31–38.

[2] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[3] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[4] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 960–964.

[5] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis." in *ACL (1)*, 2013, pp. 973–982.

[6] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 59–66.

[7] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Communication*, vol. 75, pp. 27–49, 2015.

[8] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[9] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 970–974.

[10] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, L.-P. Morency *et al.*, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[11] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *4th Audio/Visual Emotion Challenge Proc.* ACM, 2014a, pp. 65–72.

[12] H. Pérez Espinosa, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 49–55.

[13] M. Sidorov and W. Minker, "Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 81–86.

[14] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd." in *Interspeech*, 2013, pp. 847–851.

[15] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[16] J. Zinken, K. Zinken, J. C. Wilson, L. Butler, and T. Skinner, "Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression," *Psychiatry research*, vol. 179, no. 2, pp. 181–186, 2010.

[17] M. R. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 136–143.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Openearintroducing the munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.

[19] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 298–305.

[20] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *4th Audio/Visual Emotion Challenge Proc.* ACM, 2014, pp. 3–10.

[21] A. T. Beck, C. Ward, M. Mendelson *et al.*, "Beck depression inventory (bdi)," *Arch Gen Psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.

[22] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally normalized transition-based neural networks," *arXiv preprint arXiv:1603.06042*, 2016.

[23] S. Pakhomov, D. Chacon, M. Wicklund, and J. Gundel, "Computerized assessment of syntactic complexity in alzheimers disease: a case study of iris murdochs writing," *Behavior research methods*, vol. 43, no. 1, pp. 136–144, 2011.

[24] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[25] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system (facs)," *A technique for the measurement of facial action. Consulting, Palo Alto*, vol. 22, 1978.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[27] G. Saon, "Word error rate: Recent advances in conversational speech recognition," https://developer.ibm.com/watson/blog/2016/04/28/recent-advances-in-conversational-speech-recognition-2/, 2016.

[28] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.