Interspeech 2018
2-6 September 2018, Hyderabad

Mining multimodal repositories for speech affecting diseases

Joana Correia[12], *Bhiksha Raj* [1], *Isabel Trancoso* [2], *Francisco Teixeira* [2]

[1]Carnegie Mellon University, USA
[2] INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal

joanac@cs.cmu.edu

Abstract

The motivation for this work is to contribute to the collection of large in-the-wild multimodal datasets in which the speech of the subject is affected by certain medical conditions. Our mining effort is focused on video blogs (vlogs), and as a proof-of-concept we have selected three target diseases: Depression, Parkinson's disease, and cold.

Given the large scale nature of the online repositories, we take advantage of existing retrieval algorithms to narrow the pool of candidate videos for a given query related with the disease (e.g. depression vlog), and on top of that we apply several filtering techniques. These techniques explore both audio, video, text and metadata cues, in order to retrieve vlogs that include a single speaker which, at some point, admits that he/she is currently affected by a given disease. The use of straightforward NLP techniques on the automatically transcribed data showed that distinguishing between narratives of present and past experiences is harder than distinguishing between narratives of self experiences and of someone else's.

The three resulting speech datasets were tested with neural networks trained with speech data collected in controlled conditions, yielding results only slightly below the ones achieved with the original test datasets.

Index Terms: data mining, pathological speech

1. Introduction

Speech, being a complex bio-signal that is intrinsically related to human physiology and cognition, has the potential to provide a rich bio-marker for health, e.g. allowing a non-invasive route to early diagnosis and monitoring of a range of conditions including Parkinson's disease, anxiety, depression or dementia, just to name a few [1][2]. With the rise of speech related machine learning applications over the last decade, there has been a growing interest in developing speech based diagnosis-aid tools that perform non-invasive diagnosis [3][4][5][6][7].

However, one of the biggest challenges of developing computer-aided diagnosis systems based on speech is acquiring large amounts of training data. Often, the limited training data available is recorded in controlled conditions, raising concerns relating to the ecological validity of the experimental results obtained. At the same time, the cost of collecting data in controlled conditions is high, eventually prohibitive: From finding eligible and willing subjects, to assigning healthcare specialists, and guaranteeing the logistic and legal requirements for the data collection process.

Our motivation is to provide a proof-of-concept of a valid alternative to the traditional process of creating datasets. We argue that it can be achieved through mining medical data from in-the-wild, large scale, multimodal repositories. We hypothesize that this type of data exists in very large quantities, and contains highly varied examples of the effects of the diseases on the subjects speech, unbound by human experiment design. At the same time, this alternative keeps the collection cost low, in terms of time and human resources. To the extent of our knowledge, this is the first work attempting to automatically collect disease specific datasets from multimodal online repositories.

We describe the ideal video candidate for the dataset as: featuring a single subject; who is talking about himself/herself; referring to present and not past medical conditions; and includes a spoken confirmation of their diagnosis. Video blogs (vlogs), a popular category of videos which is defined as a personal video logging of any given experience, typically with little production and editing, usually contain most of the aforementioned characteristics. Therefore, we focused our mining efforts on them, (p.e. the query is *depression vlog*) in order to help exclude other video formats also related to the target disease, such as news pieces, lectures, etc. Even then, the fraction of target videos is typically less than half the total number of retrieved videos. Therefore, it is necessary to filter out the videos that do not contain first person and present experiences about the target disease. To do so, we propose doing a multimodal analysis of the video and its metadata, using mostly off-the-shelf tools in order to test the potential of our approach.

As a proof-of-concept we have selected three target diseases: Depression, Parkinson's disease, and cold. We collected and labelled a small dataset for each target disease from YouTube, building a corpus of in-the-Wild Speech Medical (WSM) data, with which we test our proposed filtering solution.

Additionally, we test state-of-the-art neural networks, trained to detect pathological speech with data collected in controlled environments, against the WSM Corpus, to highlight the differences between in-the-wild pathological speech, and pathological speech collected in controlled conditions.

This paper is organized as follows: Section 2 describes the simple retrieval process used to build this initial dataset from the online repository YouTube; Section 3 reports the process of filtering out the unwanted videos, describing the multimodal feature extraction process, and the classifiers; Their performance in detecting the target videos in the WSM dataset is presented in Section 4; Section 5 describes the models and experiments performed with data in a controlled environment, and compares them to the results obtained on WSM with the same models; Finally we draw some conclusions in Section 6.

2. The WSM Corpus

The depression, Parkinson's, and cold datasets of the WSM corpus were collected in February 2018 from the online multimodal repository YouTube. The published dates of videos ranged from January 2007 to February 2018. The language of

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references UID/CEC/50021/2013, and SFRH/BD/103402/2014.

10.21437/Interspeech.2018-1806

Table 1: *Positive class incidence per label, per disease for the WSM Corpus.*

Dataset	Vlog	1st Person	Present	Target topic	All
Depression	92.2	73.4	50.0	56.3	28.1
Parkinson's	56.3	54.7	56.3	68.8	28.1
Cold	96.9	79.7	90.6	62.5	46.9

Table 2: *Overview of the WSM Corpus.*

Dataset	Class	# Videos	Ave. duration [min]	Ave. # Words/ Video	Ave. # Words/ Min./Video	Vocab. size	Total length [min]	Total length [words]
Depression	**Positive**	18	8.85	1142.44	149.28	2130	159	20564
	Negative	40	10.44	1370.98	145.78	4321	418	54839
	Overall	58	9.95	1300.05	146.86	5096	577	75403
Parkinson's	**Positive**	18	6.73	948.50	138.78	2275	121	17073
	Negative	43	10.11	1229.19	103.63	5058	435	52855
	Overall	61	9.11	1146.36	114.00	5849	556	69928
Cold	**Positive**	30	15.96	968.23	149.77	2930	479	29047
	Negative	33	10.07	1319.61	133.61	3710	332	43547
	Overall	63	12.88	1152.29	141.30	5097	811	72594

the videos was restricted to English. The size of the WSM Corpus has been limited to approximately 60 videos per dataset, because of the need for manual labeling.

The dataset was collected by using a combination of the official YouTube API and scrapping tools to retrieve a list of results for the query ”[*target disease*] vlog”. The following information for each result (some of the items are marked as optional, if they are not required to be filled out by the uploader): video; unique identifier; title; description (optional); transcription (automatically generated for videos in English, unless provided by a user); channel identifier; playlist identifier; date published; thumbnail; video category (closed set, 14 categories, e.g. ”News”, ”Music” or ”Entertainment”); number of views; number of thumbs up; number of thumbs down; comments.

We note that the video's transcription was automatically generated by YouTube(only for videos in English), using a large scale, semi-supervised deep neural network for acoustic modeling [8], unless a human transcription is provided by a user.

Each video in WSM Corpus was hand labeled with four intermediate binary labels: 1) the video is in a vlog format; 2) the main speaker of the video talks mostly about him/herself; 3) the discourse is about present experiences or opinions; 4) the main topic of the video is related to the target disease. If all intermediate labels were positive, the video was labelled as containing in-the-wild pathological speech.

Table 1 shows the class distribution for each label, for the three datasets. Table 2 presents some statistics for each dataset, relatively to the ”all” class, namely: the average length of the videos, the average number of words in the video's transcription, the average number of words per minute; the dataset length in minutes and in words; and the total vocabulary size. These statistics are presented for each dataset, both overall and broken down for positive and negative presence of pathological speech.

3. Automatic Filtering of Videos with Pathological Speech

One of the goals of this work was to perform the distinction between videos of subjects affected by a target disease at the time of the recording and other videos possibly still related to the target disease, p.e. news pieces, presentations, classes, or forms of artistic expression. As such, we focused on extracting features that help our classifiers to automatically replicate the manual labels.

Our focus was to establish a baseline performance for this task, therefore we opted for simple straightforward techniques, both for the feature extraction stage as well as for the modeling stage. We deferred replicating state of the art techniques use to solve related problems, including multimodal emotion recognition [9][10][11], and techniques that perform the synchronization of the features across different modalities [12][13][14] as future work.

3.1. Feature extraction

The feature extraction was performed mostly using existing toolkits, in order to establish a baseline performance. From the information extracted for each video, we computed the following multimodal features:

Natural Language: Bag-of-Word (BoW) features were extracted from the video's transcription. The BoW model was used to convert a transcription in to a frequency vector of tokens in the transcriptions. In this scheme, we obtained one feature vector per transcription, in which each feature was the normalized frequency of an individual token. The length of the vector was the total size of the vocabulary of the corpus of transcriptions. This model ignored the ordering of the tokens in the transcription. In order to reduce the weight of very common words, (e.g. the, a, is in English), which carry very little meaningful information about the actual content of the document, we used the term-frequency times inverse document-frequency (tf-idf) transform.

Sentiment features were derived from the title, description, transcription and top *n* comments of the video using the Stanford Core NLP [15]. This tool is based on a Recursive Neural Tensor Network (RNTN). RNTNs take as input phrases of any length, and represent them through word vectors and a parse tree. They then compute vectors for higher nodes in the tree using the same tensor-based composition function. This RNTN was trained on a corpus of movie reviews [16], and parsed with the Stanford parser [17]. At this early stage, and given the small dataset size, we have not yet included topic modeling, neither semantic word embedding models.

Speech: We determined the number of speakers in the video, via speaker diarization to the audio component of the each video, using the LIUM toolbox [18]. The diarization process is composed of 5 steps: First music segments are removed music using Viterbi decoding; next, the signal is segmented in to speakers and background by acoustic segmentation and Hierarchical Agglomerative Clustering (HAC); then, a Gaussian Mixture Model (GMM) is trained for each cluster; the signal is then re-segmented through a Viterbi decoding; finally, the system performs another HAC, using a cross-likelihood ratio measure and the trained GMMs.

Visual: Each video was segmented into scenes, using a simple comparison between pairs of consecutive frames. Scene changes were marked when the difference exceeded a preset threshold. A random frame was selected for each resulting scene. Automatic face detection using the toolkit [19], and computation of color histograms is performed in the resulting frames.

Metadata: Features derived from the collected metadata included: a one hot vector representing the video category; the video duration; the number of views; the number of comments; the number of thumbs up; and the number of thumbs down at the time of collection.

3.2. Classifiers

We use two straightforward, well known models, to predict the intermediate labels of the videos y_i as well as the pathological speech label: Logistic regression (LR), and Support vector machines (SVMs). For the case of the SVM we train 3 distinct models with linear, polynomial of degree 3, and radial basis function (RBF) kernels.

Given the large scale nature of the online repositories, it is our hypothesis that the amount of content available per disease is much larger than the size of the desired dataset. As such, it is preferable to exclude content with a low confidence measure of containing the target disease, rather than to include it. This translates to training models that favour a high precision over a high recall.

4. Filtering Results

In order to understand the contribution of each type of feature to filter the target content, we trained a distinct classifier for each type of feature, and another one with all the features. The text component contributed with 28 features that describe the sentiment in the title, description, transcription and comments of the video; plus 5096, 5849 and 5097 BoW features, for depression, Parkinson's and cold dataset, respectively (the number differs for each dataset, based on their respective vocabulary size). The speech component contributed with a single feature describing the number of speakers in the video. The video component contributed with a 768 dimensional feature vector to describe the average color histogram of the video; plus one feature indicating the number of different faces identified in the video; and one feature indicating the number of scenes detected in the video. The metadata contributed with 19 features. By concatenating features extracted from all modalities, the final feature vectors have 5914, 6667, 5915 dimensions, for the depression, Parkinson's and cold dataset, respectively.

In total, 540 models were trained: LR, linear SVM, polynomial SVM, and SVM-RBF, for each one of the eight types of features plus one for all the features, for each of the 5 labels, per dataset in the WSM Corpus. The models were trained in a leave-one-out cross validation fashion. Given the limited amount of examples in our datasets, and the comparatively large number of features, the feature vectors were reduced in dimensionality by eliminating the features with a Pearson correlation coefficient (PCC) to the label below 0.2, thus only the features that carried some linear correlation to the label were preserved.

The results are reported in precision and recall. We consider that a good model will have a high precision measure, since the goal is to maximize the rate of true positives. At the same time, false negatives are not a major concern in this scenario: we assume that the repository being mined has a much larger number of target videos than the size of the desired dataset.

Tables 3 4 and 5 summarize the performance of the best overall model (SVM-RBF), for depression, Parkinson's, and cold, respectively. The results of the remaining models are omitted, for the sake of brevity. The cells highlighted in gray mark models which performed equal or worse than simply choosing the majority class. These models had performed poorly due to the limited amount of features available, and the excecively low dimentionality of the feature set. The best performing models for each dataset achieve a 93%, 100%, and 88% precision, and 72%, 89%, and 97% recall, for the depression, Parkinson's and cold datasets, respectively.

Table 3: *Performance of the SVM-RBF reported in precision and recall rate in detecting target content in the depression dataset of the WSM Corpus.*

Modality	Features	Label				
		Vlog	1st Person	Present	Target topic	All
Text	BoW	0.98 1.0	0.98, 1.0	0.73, 0.9375	0.89, 0.89	0.86, 0.67
	Sentiment	0.91, 1.0	0.77, 0.96	0.52, 0.66	0.52, 0.71	0.33, 0.17
Speech	#Speakers	0.91, 1.0	0.85, 0.91	0.56, 0.69	0.69, 0.94	0.0, 0.0
Video	#Faces	0.91, 1.0	0.89, 0.93	0.69, 0.75	0.72, 0.94	0.0, 0.0
	#Keyframes	0.91, 1.0	0.84, 0.96	0.56, 0.88	0.72, 0.97	0.0, 0.0
	Color hist.	0.91, 1.0	0.77, 0.98	0.69, 0.78	0.80, 0.89	0.75, 0.33
Metadata	Metadata	0.91, 1.0	0.77, 0.98	0.62, 1.0	0.60, 0.97	0.0, 0.0
All	All	0.981, 1.0	0.93, 0.96	0.83, 0.91	0.89, 0.91	**0.93, 0.72**

Table 4: *Performance of the SVM-RBF reported in precision and recall rate in detecting target content in the Parkinson's disease dataset of the WSM Corpus.*

Modality	Features	Label				
		Vlog	1st Person	Present	Target topic	All
Text	BoW	1.0, 0.86	0.74, 0.82	0.81, 1.0	0.91, 1.0	1.0, 0.89
	Sentiment	0.71, 0.71	0.69, 0.71	0.77, 0.49	0.73, 0.95	0.88, 0.39
Speech	# Speakers	0.48, 0.69	0.56, 1.0	0.57, 1.0	0.71, 1.0	0.0, 0.0
Video	# Faces	0.63, 0.94	0.58, 0.85	0.58, 0.89	0.75, 0.95	0.0, 0.0
	# Keyframes	0.55, 0.89	0.51, 0.82	0.54, 0.89	0.72, 0.91	0.0, 0.0
	Color hist.	0.76, 0.71	0.69, 0.73	0.70, 0.60	0.69, 0.95	0.0, 0.0
Metadata	Metadata	0.73, 0.77	0.49, 0.76	0.56, 0.77	0.70, 0.98	0.0, 0.0
All	all	0.97, 0.91	0.87, 0.82	0.80, 0.91	0.90, 1.0	**1.0, 0.89**

The Tables show the contribution of each type of feature to the overall performance, as well as the performance of the model in identifying each intermediate label correctly, and the final label. The type of features that has the most impact are the text features, concretely, the Bag-of-words, for every dataset, and for every label. They convey, in fact, for the Parkinson's and cold datasets, sufficient information to achieve the best performance, without any other type of feature. Overall, it is not clear which are the features, other than the bag-of-words that consistently contribute to the good performance of the models. Label 3 (Present) was the hardest label to correctly estimate, in two out of the three datasets. The results for Label 1 (Vlog) are not reported for Table 5, because the cold dataset did not contain enough negative examples to allow model training. We note that some feature types, such as the number of speakers or the number of scenes, are seldom capable of generating a good model, probably due to the limitations of the feature extraction techniques.

5. Comparing the WSM Corpus to Datasets Collected in Controlled Conditions

Neural networks trained with data collected in controlled conditions, were used to detect pathological speech in the WSM Corpus and their original test datasets. We only report results for the depression and cold corpus, since at the time of making this work we did not have a dataset for Parkinson's detection using speech.

Table 5: *Performance of the SVM-RBF reported in precision and recall rate in detecting target content in the cold dataset of the WSM Corpus.*

Modality	Features	Label				
		Vlog	1st Person	Present	Target topic	All
Text	BoW	NA	1.0, 1.0	1.0, 1.0	0.95, 1.0	0.88, 0.97
	Sentiment	NA	0.81, 1.0	0.92, 1.0	0.64, 0.85	0.64, 0.53
Speech	# Speakers	NA	0.81, 1.0	0.92, 1.0	0.63, 1.0	0.70, 0.53
Video	# Faces	NA	0.81, 1.0	0.92, 1.0	0.72, 1.0	0.71, 0.5
	# Keyframes	NA	0.85, 0.98	0.92, 1.0	0.67, 0.97	0.56, 0.67
	Color hist.	NA	0.81, 1.0	0.92, 1.0	0.65, 0.93	0.60, 0.5
Metadata	Metadata	NA	0.81, 1.0	0.92, 1.0	0.65, 0.97	0.57, 0.40
All	All	NA	1.0, 1.0	1.0, 1.0	0.95, 1.0	**0.88, 0.97**

5.1. Controlled Conditions Datasets

5.1.1. Depression

The depression subset of the Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ) [20] is an audio-visual database of clinical interviews. It consists of 189 sessions ranging between 7 and 33 minutes, 106 of which are present in the training set, and 34 in the development set. For each of these sessions a score is provided in the PHQ-8 [21] scale as a measure of depression. Of the 106 participants in the training partition, 30 are considered to be depressed. In the development set, 34 subjects are classified as depressed [22].

5.1.2. Cold

The Upper Respiratory Tract Infection Corpus (URTIC) [2] is a dataset provided by the Institute of Safety Technology of the University of Wuppertal, Germany, for the Interspeech 2017 ComParE Challenge. It contains recordings of spontaneous and scripted speech. The training and development partitions comprised 210 subjects each, but only 37 had a cold. The two partitions include 9,505 and 9,565 chunks of 3 to 10 seconds, respectively [2].

5.2. Feature Extraction

For depression and cold datasets, we used extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features, a set of 88 acoustic features designed to serve as a standard for paralinguistic analysis [23].

The DAIC and URTIC corpus were already segmented. The WSM Corpus underwent automatic diarization, prior to feature extraction, using LIUM, to eliminate silent segments, and divide the speech signal into inter-pausal units. The segments that did not belong to the main speaker were discarded. The minimum segment length was set to $200ms$.

5.3. Model

We follow a simple neural network structure for the model. It consists of three layers: an input layer with 120 units, a hidden layer with 50 units, and an output layer with one unit. The first two layers share the same structure, first a Fully Connected (FC) layer, followed by a Batch Normalization (BN) layer, and an Activation layer with Rectified Linear Units (ReLUs). The output layer is characterized by a FC layer with a sigmoid activation. During training, Dropout layers are also inserted before the second and third FC layers. Both the Dropout and the BN layers in the network help prevent the model from overfitting [24] [25]. These forms of regularization are important in this case, due to the limited size of the training data.

Before training the network, the training set is zero-centered and normalized by its standard deviation. The values of the mean and standard deviation of this set are later used to zero-center and normalize the development set.

The model was implemented in Keras[26], and was trained with RMSProp, using the default values of this algorithm together with a learning rate of 0.02 and 100 epochs. To determine the best dropout probabilities for each dropout layer, a random search was conducted yielding the following values: 0.092 and 0.209 in the depression model; and 0.3746 and 0.5838 for the second cold model, for the first and second dropout layers, respectively.

To compensate for the unbalanced labels on the training partitions of the depression and cold datasets, we attribute different weight to samples of the positive and negative class: 0.8/0.2 for depression, and 0.9/0.1 for cold.

Table 6: *Comparison of the performance in UAR of the Neural Networks to detect pathological speech in datasets collected in controlled environments versus data collected in-the-wild.*

Voice affecting disiease	**Controlled Conditions Dataset**		**WSM Corpus**	
	Train (segment level)	**Development (segment level)**	**Development (segment level)**	**Development (speaker level)**
Depression	60.59	60.57	54.79	61.94
Cold	59.95	66.92	53.11	54.81

5.4. Results

The performances in precision and recall of the neural networks against the WSM Corpus versus existing datasets of data collected in controlled conditions are summarized in Table 6. As expected, given the greater variability in recording conditions (p.e. microphones, noise), the performances of the networks when faces with in the wild data decrease when compared to data collected in controlled conditions. However, it is possible to improve the classification at speaker level, versus at segment level by aggregating the segments for each speaker, as the last column of Table 6 shows, particularly in the case of depression. The subject level prediction, obtained by computing a weighted average of the segment level predictions, in which the weighting term is given by the segment length.

We hypothesize that an additional justification for the performance drop is the greater variability in the speech alterations of e speakers in the in-the-wild datasets, given that their discourse is not bounded, as could be the case in a controlled environment, thus facing the networks with unseen speech alterations.

6. Conclusion

This work established a baseline for collecting disease specific datasets of in-the-wild data, containing instances of speech affecting diseases, based on mining multimodal online repositories. We demonstrated the viability of this process for three diseases: depression, Parkinson's, and cold, which leads us to believe that the process is generalizable to collect datasets for any disease. Given its modular nature, each component of the system, can be individually improved.

The best performing models achieved a precision of 93%, 100%, and 88%, and a recall of 72%, 89%, and 97%, for the dataset of depression, Parkinson's, and cold respectively, in the task of filtering videos containing speech affecting diseases.

At the same time, we compared the WSM Corpus to datasets of data collected in controlled conditions. The performance in precision and recall of the existing models decreased when faced with in-the-wild data, compared to data collected in controlled conditions. We hypothesize this is due to a greater variability in recording conditions (p.e. microphone, noise) and in the effects of speech altering diseases in the subjects' speech.

For future work, we will focus on three problems: collecting and making available larger datasets of several speech affecting diseases, thus increasing the dataset resources available for medical applications; improving the performance of each individual module of our proposed system, replacing them with disease specific tools; and most importantly, moving towards a completely unsupervised data collection system, by dropping the label requirements during the training stage.

7. References

[1] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Hönig, and E. Nöth, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1820–1828, 2015.

[2] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017.

[3] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013.

[4] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, "On automatic diagnosis of alzheimers disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[5] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, vol. 1. IEEE, 2002, pp. 182–183.

[6] 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.

[7] J. Correia, I. Trancoso, and B. Raj, "Detecting psychological distress in adults through transcriptions of clinical interviews," in *International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 162–171.

[8] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 368–373.

[9] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[10] D. Palaz, M. Magimai.-Doss, and R. Collobert, "Analysis of cnn-based speech recognition system using raw speech as input," Idiap, Tech. Rep., 2015.

[11] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[12] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," in *Referring Phenomena in a Multimedia Context and their Computational Treatment*. Association for Computational Linguistics, 1997, pp. 1–13.

[13] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[14] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "Identifying human behaviors using synchronized audio-visual cues," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 54–66, 2017.

[15] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[16] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 115–124.

[17] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st annual meeting of the association for computational linguistics*, 2003.

[18] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Interspeech*, 2013.

[19] A. Geitgey, "Facerecog," https://github.com/ageitgey/face_recognition, 2017.

[20] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*. Citeseer, 2014, pp. 3123–3128.

[21] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J Affect Disord*, vol. 114, no. 1-3, pp. 163–173, Apr 2009.

[22] M. F. Valstar, J. Gratch, B. W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 - depression, mood, and emotion recognition workshop and challenge," *CoRR*, vol. abs/1605.01600, 2016. [Online]. Available: http://arxiv.org/abs/1605.01600

[23] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andr, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 4 2016, open access.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *CoRR*, vol. abs/1502.03167, 2015.

[25] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[26] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.