INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Detecting Depression with Word-Level Multimodal Fusion

Morteza Rohanian, Julian Hough, Matthew Purver

School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
{m.rohanian, j.hough, m.purver}@qmul.ac.uk

Abstract

Semi-structured clinical interviews are frequently used diagnostic tools for identifying depression during an assessment phase. In addition to the lexical content of a patient's responses, multimodal cues concurrent with the responses are indicators of their motor and cognitive state, including those derivable from their voice quality and gestural behaviour. In this paper, we use information from different modalities in order to train a classifier capable of detecting the binary state of a subject (clinically depressed or not), as well as the level of their depression. We propose a model that is able to perform modality fusion incrementally after each word in an utterance using a time-dependent recurrent approach in a deep learning set-up. To mitigate noisy modalities, we utilize fusion gates that control the degree to which the audio or visual modality contributes to the final prediction. Our results show the effectiveness of word-level multimodal fusion, achieving state-of-the-art results in depression detection and outperforming early feature-level and late fusion techniques.

Index Terms: depression, recurrent neural networks, computational paralinguistics, modality fusion, gestural behaviour, lexical content

1. Introduction

The automatic diagnosis of depression has gained popularity in recent years: depression has a high degree of public prevalence and is one of the most serious forms of disability worldwide [1]. Diagnosis and assessment for depression is generally based around the judgement of clinicians, and commonly uses semi-structured interviews, guided by predetermined sets of topics, in a clinical set-up.

Depression causes cognitive and motor changes that affect speech production: reduction in verbal activity productivity, prosodic speech irregularities and monotonous speech have all been shown to be symptomatic of depression [2]. Depressed patients' spectral-based features have been observed as changing noticeably in depressive states [3]. Their affective state is also influenced by the condition, indicated through prosodic features [4]. However despite several factors being mildly predictive of a depressive state, it has been claimed that because of the innate differences in speaking manner, no single feature on its own has enough discriminatory power as an indicator of depression [5].

Paralinguistic nonverbal cues have been used as depression markers in clinical sessions. Depressed patients exhibit less facial expressivity [6] and less frequent mouth movement [7]. They are more likely to have impaired attention and keep mutual gaze less frequently [8], turn away their gaze and turn their heads down [6]. In addition to nonverbal behavior, linguistic analysis displays important depression indicators. The lexical content of a patient's utterances in clinical interviews has been shown to be effective in detecting depression [9]. Considering the broad clinical outline of depression, it seems that there are significant benefits to be gained from a *multimodal* approach to detecting depression, integrating features from sets of verbal and nonverbal channels of communication.

2. Previous work on depression and cognitive state detection

Recent experimental work has explored the automatic analysis of depression from multimodal data. There has been work on building systems that classify severity of depression using a wide range of multimodal features. Publicly available multimodal depression datasets, which are collections of clinical interviews, have provided an opportunity to explore a range of experiments on detecting depression. Most current approaches use either *early* feature-level fusion whereby features from the different modalities are combined into a new feature set for classification, or *late* prediction-based fusion whereby separate classifiers are trained on each modality to predict the depression state and the the output of those classifiers are combined into a single prediction. Meng et al. use Partial Least Square (PLS) regression for predicting depression based on each modality and apply a late fusion method for the final prediction [10]. Yu et al. propose a multimodal Hidden Conditional Random Fields (HCRF) model considering question and response pairs [11]. Along the same line, Gong et al. combine topic modeling of question/answer of the interviews with multi-modal text, audio, and video features to predict depression levels [12]. Yang et al. use manually selected features as input into a Deep Convolutional Neural Network (DCNN). The learned features are fed to a Deep Neural Network (DNN) to predict the severity of depression [13].

In terms of the communicative features which aid depression detection, lexical features from the interviewer's utterances are shown to be an informative feature for depression in a multimodal classification task with a staircase Gaussian approach [14]. There has also been work on modelling unimodal sequential input for depression detection. Ma et al. propose an audio based method for depression classification using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for a higher level audio representation [15]. Sun et al. present a unimodal random forest method based on the question/answer characteristics of the interview sessions [16].

Nasir et al. consider the temporal nature of audio/visual modalities using a window-based representation of the features instead of the more common approach of frame-level analysis [17]. Utilizing complementary information from text and audio features, Alhanai et al. proposed a model in which two LSTM branches, one per modality, are integrated via a feed-forward network [18]. However, while this work tries to predict depression based on late or early fusion methods [10, 12] or the sequential nature of their inputs [17, 18], learning the

http://dx.doi.org/10.21437/Interspeech.2019-2283

time-dependent relationships between language, visual and audio features in detecting depression is still unexplored.

In other related tasks using multimodal fusion to predict a cognitive state, there has been work on combining temporal information from two or more modalities in a recurrent approach in audio/visual emotion classification [19] and image captioning tasks [20]. This work demonstrated the ability to learn complicated decision boundaries that other models with different fusion methods have difficulty handling [21]. One major problem these models have is dealing with the different predictive power of each modality and their different levels and types of noise. Adding gating mechanisms has been shown to be effective in dealing with the level of contribution of each modality to the final prediction in different multimodal tasks [22, 23].

Our approach is motivated by some of the recent efforts in multimodal fusion for classifying cognitive states to capture the interaction between modalities in detecting depression and maximise the use and combination of each modality. In this paper we propose a *word-level* multimodal fusion with a simple gating mechanism in a time-dependent recurrent framework, and compare it with early and late fusion techniques.

3. Proposed Approach: Word-level multimodal fusion with gating

To predict the severity of depression based on learning multimodal representations, we explore three techniques for fusion: early, late and a model-based approach in which optimal fusion is learned using a neural network. We explore the use of a gating mechanism to learn how best to filter the visual and auditory modalities' effect on lexical information.

3.1. Pre-processing: Forced Alignment for word timings

An essential part of multimodal representation is to model the inter-modality dynamics: to properly learn the time-dependent interactions between language, visual and audio features and integrate them using timestamps. While in a live system we would use time-stamps from a speech recognizer, for this proof-of-concept study we perform offline forced alignment between text, audio and visual features to get the precise time-stamp of every uttered word. At every time-step, we align words with their matching audio time interval using the Penn Phonetics Lab Forced Aligner (P2FA) [24]. P2FA is a tool that can be applied to align transcriptions to audio files, phoneme by phoneme. Upon manual inspection the forced alignment was performed with high enough accuracy for the fusion study in this paper.

3.2. Gating Mechanism

Data from the three modalities have different effects on the final output and it is important to consider the amount of noise when aggregating them into a representation. Since learned representation for the text can be undermined by corresponding visual and audio modalities, we need to alleviate the effects of noise and overlap during multimodal fusion. One way to overcome this problem is to go beyond naive concatenation of vectors representing either the features themselves, or predictions derived from them, and control the degree to which, the audio and visual data contribute to the final prediction using a simple gating mechanism.

We utilize feed-forward highway layers [25], with gating units which learn to regulate information flow through the network by weighting visual and audio inputs at each time-step.

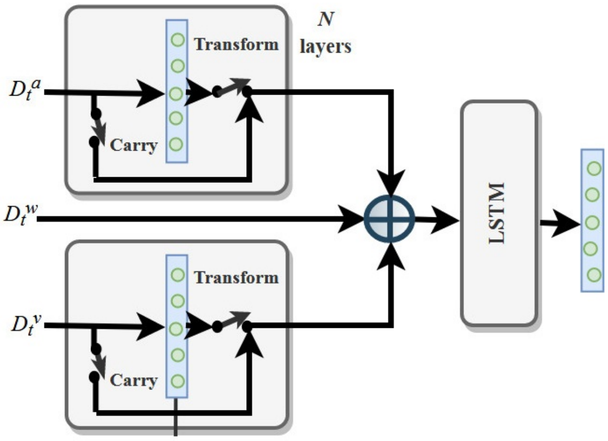

Figure 1: *Word-level multimodal fusion with gating.*

Each highway layer comprises two non-linear transforms: a Carry (Cr) and a Transform (Tr) gate which define the degree to which the output is created by transforming the input and carrying it (how much information should move forward or be changed in successive training epochs). Each layer controls its input vector D_t using the gates and a feed-forward layer H:

$$y = Tr \cdot H + Cr \cdot D_t \quad (1)$$

where Cr is simply defined as $1 - Tr$, giving:

$$y = Tr \cdot H + (1 - Tr) \cdot D_t \quad (2)$$

The transform gate Tr is defined as $\sigma(W_{Tr} D_t + b_{Tr})$, where W_{Tr} is the weight matrix and b_{Tr} the bias vector for the gates. Based on the outputs of the transform gates, highway layers can change their performance from layers made of multiple units to layers which only pass their inputs through. As inspired by [25] and to help overcome long-term dependencies earlier in learning, we initialize b_{Tr} with a negative value (biased towards the Carry gate). We use a block of stacked highway layers.

3.3. Model Architecture

We set our model up to learn the most useful interactions between modalities for predicting depression. To achieve this, feature vectors from the three modalities are concatenated to create the input D_t to a word-level LSTM at each time-step t. The overall architecture of our LSTM with Gating model is shown in Figure 1. The gating mechanism is first applied to the audio and visual feature input vectors D_t^a and D_t^v which are passed through N highway layers (where the best value N is determined from optimizing on heldout data) before being concatenated with the current word embedding D_t^w to form the input vector to the LSTM network. After training our LSTM with gating, the resulting Mean Absolute Error (MAE) loss is used as the signal for training our highway layers, employing the REINFORCE rule [26] in a similar way to [27].

Fusion comparison. In addition to testing the effect of using full multi-modality as described compared to combinations of two modalities and single modalities, and also investigating the effect of the gating mechanism, we also compare our model-based fusion technique to two commonly used fusion techniques: early (i.e., feature-based) and late (i.e., decision-based) fusion. In early fusion we integrate features right after extraction (by concatenating them), passing the concatenated feature vector as input into the LSTM. The late fusion classifier obtains unimodal decision values from three different LSTMs, one for each modality, and then combines their decisions using a weighting mechanism for the final prediction.

4. Experiments

Data. We experiment with datasets from the publicly available Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) with audio, text transcripts and visual features [28]. The DAIC dataset contains clinical interviews, conducted by an animated virtual agent. The training, development, and test sets contain 107, 35, and 47 subjects and the state of the subjects is evaluated based on the PHQ-8 metric [29]. The PHQ-8 assessment rates the severity of symptoms detected in depression, like anxiety, insomnia and agitation to assign a score to a patient based on their level of depression. In addition to binary state of subjects, we predict different degrees of depression at the subject level on the designated test set. The level of depression ranges from 0 to 24 with the range 0-4 regarded as not depressed, 5-9, 10-14 and 15-19 as moderate and 20+ as severe.

4.1. Multimodal features

Lexical Features from Text A Pre-trained GloVe model [30] with a 300-dimensional embedding space was used to extract the lexical feature representations from the transcript. We convert the sequences of responses into word vectors, without considering the queries that led to the responses.[1]

Audio Features A set of audio features are extracted using the COVAREP acoustic analysis framework software [31]. The features include 12 Mel-frequency cepstral coefficients (MFCCs), voiced/unvoiced segmentation and pitch tracking [32], peak slope and maximal dispersion quotients, glottal source parameters (using glottal inverse filtering of GCI synchronous IAIF) [33] and shape parameter of the Liljencrants-Fant model of glottal pulse dynamics [34]. These audio features are extracted based on various attributes of human voice that have been shown to be helpful in detecting depression [5]. Since words are the fundamental units of the input in our models, the interval duration of each word is used as a time interval for capturing these features for each input step. The values for each 10ms frame are averaged to make a single vector for the current word's duration.

Visual Features The visual features are frame-level (20ms window, 10ms shift), provided with the DAIC dataset. They are extracted using the library OpenFace [35] which includes estimates of head position, head rotation, 68 facial landmark locations, gaze tracking, facial action units (FAUs) and HOG features [36]. As with the audio, the average of the frame-level features of the interval duration of each word are used as the visual modality information.

4.2. Implementation and Metrics

All of the experiments are performed without conditioning on speaker identity. The layer sizes and the learning rates are determined using grid search on validation data. The *N* for Highway networks is an additional hyperparameter required over standard recurrent deep approaches, and 3 was found to be the optimal value. The LSTM models have 128 hidden nodes and are trained using ADAM [37] with learning rate 0.0001. The Mean Absolute Error (MAE) from the ground-truth PHQ-8 assessment scores for each subject is used as the loss function.

For binary classification of depression, we report precision and F1 score and for the PHQ-8 numeric rating accuracy we report the MAE and Root Mean Square Error (RMSE).

4.3. Baseline Models

We compare the performance of our models to the following four models that use the DAIC dataset whose approaches are related to our work: (i) the DAIC baseline with an ensemble of features in a Support Vector Machine (SVM) model which was provided with the dataset [28]; (ii) Gong et al. which uses an ensemble of features with an approach based on topic-modeling [12], (iii) Alhanai et al.'s alternative deep learning model which uses two LSTMs (audio-based and text-based) and a final feed-forward network to model sequences of interactions for detecting depression [18]; (iv) Williamson et al. which performs topic-dependent fusion scoring on text, audio and video [14].

5. Results

Table 1: *Result of the depression classification experiments with our models against state-of-the-art competitors*

Model	**Features**	**F1**	**Prec.**	**MAE**	**RMSE**
Baselines					
DAIC Baseline [28]	Audio+Visual	-	-	5.66	7.05
Gong et al. [12]	Text+Audio+Visual	0.60	-	3.96	**4.99**
Alhanai et al. [18]	Text	0.66	0.70	5.09	6.11
Alhanai et al. [18]	Text+Audio	0.75	0.72	5.02	6.04
Williamson et al. [14]	Text	0.67	0.74	3.82	5.06
Williamson et al. [14]	Text+Audio+Visual	0.70	0.78	3.84	5.23
Our Models					
LSTM	Text	0.69	0.68	4.98	6.05
LSTM	Text+Audio	0.67	0.68	5.18	6.40
LSTM	Text+Audio+Visual	0.67	0.63	5.29	6.68
LSTM with Gating	Text+Audio	0.80	0.78	3.66	5.14
LSTM with Gating	Text+Audio+Visual	**0.81**	**0.80**	**3.61**	**4.99**

In Table 1, we present our proposed word-level fusion model's performance against that of baselines and previous state-of-the-art models on depression detection on the provided test set. For detecting depression, our proposed word-level fusion LSTM model with gating achieves an F1 score of 0.81 and MAE of 3.61, outperforming all the baselines. The overall results support our assumption that a model with gating mechanisms can mitigate the errors and noise of individual modalities most effectively.

The LSTM model with gating outperforms other multimodal and single modality depression detection models in both binary and multi-class classification tasks. There is a significant performance boost by integrating textual and audio modalities with gating over not using it (F1 0.80 vs. 0.69; MAE 3.66 vs. 4.98). Adding visual features improves the performance despite the fact that word-alignment models cannot be easily used to combine frame-level visual information due to the fact the relatively slow frame rate from the visual information does not allow consistent overlap with the input word's duration (F1 0.81 vs. 0.69; MAE 3.61 vs. 3.66). The text features are highly informative for depression classification on their own, and without the appropriate fusion techniques the performance level can in fact decrease: integrating other modalities without gating control led to a slightly worse performance in our experiments (F1 scores 0.67 vs. 0.69; MAE 5.29 vs. 4.98).

In terms of our competitor baselines, while [18] and [14]'s multimodal classifiers performed better than all the unimodal

[1] Note this differs to [14] who found the interviewer's questions to contain highly predictive features.

Table 2: *Depression classification results using Unimodal features*

Model	F1	Prec.	MAE	RMSE
LSTM with Lexical Features	0.69	0.68	4.98	6.05
LSTM with Audio Features	0.66	0.71	5.21	6.44
LSTM with Visual Features	0.59	0.63	5.38	6.72

Table 3: *Depression classification results of systems with different fusion techniques*

Fusion Method	F1	Prec.	MAE	RMSE
Early Fusion	0.67	0.63	5.29	6.68
Late Fusion with Weighting	0.70	0.78	3.92	5.86
Model-Based Fusion	0.81	0.80	3.61	4.99

models, showing some useful fusion, we note that they both utilized utterance-level fusion and ignored the time-scale associations, meaning that these models may not function word-by-word incrementally. For integration into any live system, we suggest incremental processing is vital. Furthermore, our model outperforms models without utilizing the topic/context of questions and sequences of responses [12, 14] and the model with word-level audio features achieves better F1 and MAE performance in comparison to Alhanai et al. [18] that uses set of higher-order statistics as audio features for each individual's response. This indicates the potential advantages of an incremental word-level structure over employing global information across different time scales, without needing look-ahead for utterance-global or dialogue-global features. The model we proposed, utilizing sequence of utterances and trying to capture important temporal interactions, without conditioning on the topic of the query, performs better than [14]'s state-of-the-art baselines with context/topic modeling (F1 0.81 vs. 0.70 and MAE 3.61 vs. 3.84).

5.1. Fusion Analysis

Text is the most influential modality in detecting depression in a word-level structure in this dataset. From Table 2, we can see the performance of our LSTM models across modalities. Using only the text modality gives a better depression prediction than utilizing unimodal audio and visual modalities sequentially. Adding modalities to the LSTM with text without gating does not lead to improvement. Utilizing more modalities even results in worse performance in both MAE and F1 compared to unimodal LSTM with lexical features alone (Table 1). The audio and visual modalities can negatively impact the model's performance if word-level multimodal fusion is not controlled.

Our models, integrating multimodal features for each word, show improvement over Alhanai et al. [18] which attempted to find optimal input parameters for each modality, showing the potential advantages of a word-level time-dependent approach with effective fusion. When we employ gating, Table 1 indicates that more input modalities leads to better results in both F1 and MAE. We assume that the LSTM with gating succeeds in dealing with features in different contexts conveying different information at different rates and contributing different parts of the overall representation in the network. While the lexical content of the subjects' responses is clearly a strong indicator of depression in this dataset, the acoustic quality of each word is also indicative of depression, and visual information based on the bodily movement of the subject concurrent with their words also helps depression classification, albeit less markedly. While our simple technique of capturing information over word durations works well here, in future work we will explore more principled ways of capturing gesture/bodily movement data before its combination with lexical and acoustic data.

In terms of fusion techniques, the results in Table 3 show the model-based fusion method, designed to perform multimodal fusion within the network's architecture, obtains the highest performance. It benefits from observing temporal multimodal information and the ability to train both the multimodal representation and the fusion component simultaneously. The late fusion model performs better than the early fusion method (F1 0.70 vs. 0.67 and MAE 3.92 vs. 5.29) with the precision close to the model-based methods (Prec. 0.78 vs. 0.80). Late fusion approaches have the advantage of interpretability in terms of showing which modality is given the highest weight in the input, but they do not make use of the possible dependencies between modalities in real-time communication. Early fusion only needs one model for all modalities, making it the easiest and fastest method for training, however the network is not learning from the large heterogeneous input vector as effectively as the model-based version.

6. Conclusion

We have presented a model that learns the indicators of depression from audio and visual modalities as well as lexical information in transcript texts. We utilized word-level multimodal fusion with feed-forward highway layers as a gating mechanism. Our principal motivation is to capture inter-modal dynamics in a joint multimodal representation. Our model outperforms the state-of-the-art methods in both binary and numeric depression classification tasks.

In future work we intend to analyze the interactions between different modalities as the predictors of depression as they occur in real time. Monitoring the multimodal fusion after each word could help highlight informative moments that contribute more to the prediction of depression, which could in turn have several clinical applications for psychiatric practitioners in helping further understand symptoms of depression during interaction. Furthermore, we intend to undertake a more principled approach to the visual modality in terms of extracting bodily action sequences from motion capture data, which in turn interact with the verbal behaviour to give multimodal meaning.

7. Acknowledgements

We thank the reviewers for their helpful comments. Purver was partly supported by the European Unions Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors views and the Commission is not responsible for any use that may be made of the information it contains.

8. References

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Medicine*, vol. 3, no. 11, p. e442, 2006.

[2] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depres-

sion," *American Journal of Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

[3] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[4] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[5] 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.

[6] J. E. Perez and R. E. Riggio, "Nonverbal social skills and psychopathology," *Nonverbal Behavior in Clinical Settings*, pp. 17–44, 2003.

[7] J. T. M. Schelde, "Major depression: Behavioral markers of depression and recovery," *The Journal of Nervous and Mental Disease*, vol. 186, no. 3, pp. 133–140, 1998.

[8] P. Waxer, "Nonverbal cues for depression." *Journal of Abnormal Psychology*, vol. 83, no. 3, p. 319, 1974.

[9] C. Segrin, "Social skills deficits associated with depression," *Clinical Psychology Review*, vol. 20, no. 3, pp. 379–403, 2000.

[10] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 21–30.

[11] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Semdial 2013 DialDam: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, 2013, pp. 160–169.

[12] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 69–76.

[13] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 53–59.

[14] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 11–18.

[15] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 35–42.

[16] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 61–68.

[17] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 43–50.

[18] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, 2018, pp. 1716–1720.

[19] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2362–2365.

[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[21] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[22] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 163–171.

[23] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network," *arXiv preprint arXiv:1807.06233*, 2018.

[24] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus."

[25] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[26] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[27] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[28] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[29] S. Gilbody, D. Richards, S. Brealey, and C. Hewitt, "Screening for depression in medical settings with the patient health questionnaire (phq): a diagnostic meta-analysis," *Journal of General Internal Medicine*, vol. 22, no. 11, pp. 1596–1602, 2007.

[30] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

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

[32] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[33] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012.

[34] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Sweden, Tech. Rep. 4, 1985.

[35] 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.

[36] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1491–1498.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.