Estimating Severity of Depression from Acoustic Features and Embeddings of Natural Speech

Sri Harsha Dumpala
Vector Institute and
Faculty of Computer Science,
Dalhousie University
Canada
sriharsha.d@dal.ca

Sheri Rempel
Nova Scotia Health, Halifax
Halifax, Canada
sheri.rempel@nshealth.ca

Katerina Dikaios
Department of Psychiatry,
Dalhousie University and
Nova Scotia Health, Halifax
Halifax, Canada
katerina.dikaios@dal.ca

Mehri Sajjadian
Department of Psychiatry,
Dalhousie University and
Nova Scotia Health, Halifax
Halifax, Canada
mehri.sajjadian@dal.ca

Rudolf Uher
Department of Psychiatry,
Dalhousie University and
Nova Scotia Health
Halifax, Canada
uher@dal.ca

Sageev Oore
Faculty of Computer Science
Dalhousie University and
Vector Institute
Canada
sageev@dal.ca

***Abstract*—Major depressive disorder, referred to as depression, is a leading cause of disability, absence from work, and premature death. Automatic assessment of depression from speech is a critical step towards improving diagnosis and treatment of depression. Previous works on depression assessment from speech considered various acoustic features extracted from speech to estimate depression severity. But performance of these approaches is not at clinical standards, and thus requires further improvement. In this work, we examine two novel approaches for improving depression severity estimation from short audio recordings of speech. Specifically, in audio recordings of a narrative by individuals diagnosed with major depressive disorder, we analyze spectral-based and excitation source-based features extracted from speech, and significance of sentiment and emotion classification in estimation of depression severity. Initial results indicate synchrony between depression scores and the sentiment and emotion labels. We propose the use of sentiment and emotion based embeddings obtained using machine learning techniques in estimation of depression severity. We also propose use of multi-task training to better estimate depression severity. We show that the proposed approaches provide additive improvements in the estimation of depression severity.**

***Index Terms*—Depression severity, speech, prosodic, spectral, multi-task learning.**

I. Introduction

Major depressive disorder (MDD), simply referred to as depression, is the leading cause of physical and mental disability worldwide [1]. Accurate determination of depression severity and its change is key to selecting effective treatment [2]. If left untreated, depression can lead to adverse outcomes including suicide [3]. Improved automatic prediction of depression severity scores could help substantially reduce its negative impact.

Paper presented at ICASSP 2021.

Previous studies have shown that speech contains important cues for detecting depression [4]–[8]. Initial studies on depression showed that the percentage pause time, and acoustic parameters extracted from fundamental frequency (F_0) contour are correlated with depression severity [4]–[6] . Further, vocal-source-based features such as jitter, shimmer and dynamics of F_0 were found to be useful bio-markers of depression [7], [8]. Based on these speech cues, machine learning techniques were proposed for detecting depression symptoms [9]–[13]. Gaussian mixture models based on prosodic features such as F_0, jitter, shimmer, and spectral features such as formants, mel-frequency cepstral coefficients (MFCCs) were initially used to detect depression from speech [9], [10]. Spectral and prosodic features along with their statistics extracted using openSMILE toolkit [14] were used to train support vector machines and random forest models for depression detection [11], [12]. Further, pre-trained deep convolutional neural networks (CNNs) were considered to extract acoustic embeddings for depression detection [13]. More recently, CNNs were proposed for detecting depression from speech [15], [16].

In this study, we propose additional classification tasks within a multi-task learning framework [17]–[19] to improve CNN performance for estimating depression severity from speech. Our CNN layers have multi-sized kernels (previously used in text-based sentiment classification [20], [21]) to better capture inter-relations in the acoustic features at various resolutions.

Sentiment of speech and emotive state of the speaker are shown to provide important cues for detecting depression [8], [22]–[25]. Cheng et al [23] review importance of emotion recognition from speech for depression detection. Emotive information extracted from speech, text, and facial expressions proved effective in depression detection [22], [24]. Previ-

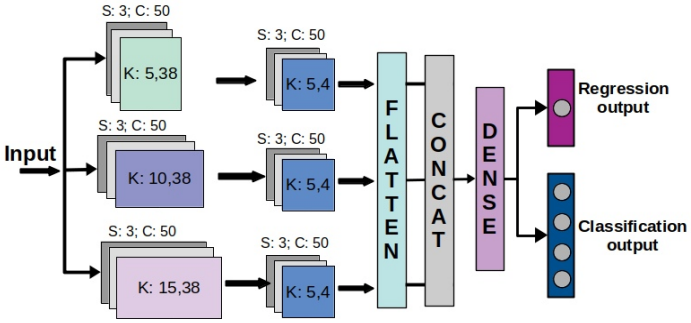

Fig. 1. Block diagram of proposed multi-task CNN for estimation of depression severity. Here, K, S, C refer to kernel size, stride and number of channels, respectively. Input can be spectral or spectral+prosodic.

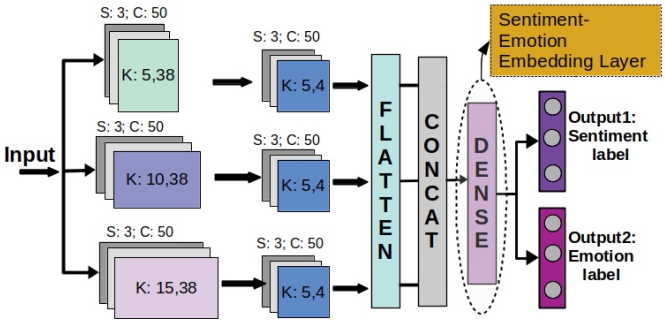

Fig. 2. Block diagram of proposed multi-task CNN for sentiment and emotion classification (*MT-CNN-SE*) to obtain the sentiment-emotion embeddings.

ous studies have identified increased expression of negative thoughts and feelings in people suffering with depression [8]. Further, text-based sentiment analysis also helped in detection of depression [25]. In this paper, we consider speech-based sentiment and emotion embeddings for estimating severity of depression. We also propose the use of combined sentiment and emotion embeddings, generated by training a multi-task CNN on sentiment and emotion classification tasks, for improved estimation of depression severity.

The rest of the paper is organized as follows. Section II explains the feature extraction, neural network architectures proposed for estimating depression severity, and for extracting sentiment and emotion based embeddings. Section III provides dataset details. Section IV discusses experimental results, with a summary of the paper in Section V.

II. Proposed Approach

We now provide details about acoustic features and network architectures we use for estimating depression severity by considering only raw speech audio recordings (i.e. no text) collected from subjects.

A. Acoustic Features

We consider spectral (Mel filter bank (MFB)) features, which predominantly carry the acoustic information, and excitation source based features (aperiodic and F_0), which carry the prosodic information. The WORLD vocoder system [26] is used to extract these features from the raw speech signals sampled at 16KHz: 36-dimensional MFBs, 1-dimensional aperiodic (AP) and 1-dimensional F_0 values are extracted from overlapping speech signal windows of length 50-msec spaced every 30-msec. Mean-variance normalization of the feature vector is performed prior to training and testing machine learning models.

B. Multi-task-CNNs

Multi-task-CNNs (*MT-CNN*s), as shown in Figure 1, are used to predict depression severity scores from extracted features. The first convolutional layer consists of 3 different kernels with sizes 5, 10 and 15, respectively. Each kernel consists of 50 channels. In the second convolutional layer, all kernels are of size 3 with 50 channels in each kernel. Outputs from each kernel of the second convolutional layer are flattened and then concatenated before passing through a dense or fully connected (FC) layer with 150 units. This FC layer is also used to obtain depression-specific embeddings representing the input sequence of acoustic feature vectors. The output layer is divided into two parts: one for regression (1 unit) and the other for the auxiliary classification task (4 units). For the regression task, the objective is to estimate the depression severity score from the given acoustic features. In the classification task, the range of standard depression severity scores (described in IV) are divided into 4 parts i.e., $0-4$, $5-8$, $9-13$, $14-21$. The variation in the ranges reflects the non-uniform empirical distribution of depression severity scores. As a baseline, we also trained a single task CNN network whose final output layer contains only the regression value.

To analyze the importance of spectral and prosodic features in estimating depression severity, two different feature sets are considered: spectral (36-dimensional) and spectral+prosodic (38-dimensional: spectral + F_0 + aperiodic) features. Separate MT-CNN networks are trained on spectral, and spectral+prosodic features, respectively.

C. Sentiment and Emotion Embeddings

To extract the sentiment and the emotion embeddings from audio features, we use the MT-CNN-SE as shown in Figure 2. Here we see that the MT-CNN-SE for extracting sentiment and emotion embeddings is similar to that in Figure 1, with few modifications. In the first convolutional layer, one dimension of the kernel size is the same as the input feature vector size (i.e., 36 for spectral and 38 for spectral+prosodic). The second convolutional layer kernel is 1-dimensional. The subsequent dense or FC layer, which we refer to as the Sentiment-Emotion Embedding Layer, consists of 100 units. The final layer consists of two parts: the first part consists of 3 units for predicting sentiment label (corresponding to $\{$negative, neutral, positive$\}$ sentiments); the second part consists of 3 units for predicting emotion label (corresponding to $\{$ anger or sad, neutral, happy$\}$ emotions). To obtain sentiment-specific or emotion-specific labels, the output layer will consist of only the corresponding part.

TABLE I
RMSE VALUES OF DEPRESSION SEVERITY SCORES USING SPECTRAL AND PROSODIC FEATURES.

Model	Feature	RMSE
Lin.-Reg.	spectral	8.72
	prosodic	9.46
	spectral+prosodic	8.58
SVM-Reg.	Spectral	8.85
	Prosodic	9.51
	Spectral + Prosodic	8.75
CNN-Reg.	Spectral	8.61
	Prosodic	9.86
	Spectral + Prosodic	8.42
Multi-task (MT-CNN)	Spectral	7.48
	Prosodic	9.05
	Spectral + Prosodic	7.36

D. Combined Embeddings

The depression embeddings (obtained from the MT-CNN in Figure 1 trained on depression regression and classification tasks) and the sentiment-emotion embeddings (obtained from the MT-CNN-SE in Figure 2 trained on sentiment and emotion classification tasks) are concatenated to obtain the combined embeddings. Speech samples are passed through the trained MT-CNN-SE to obtain the corresponding embeddings. The dimension of the combined embeddings is 250 (150-dimensional depression-specific embeddings, and 100-dimensional sentiment-emotion embeddings). These embeddings are then used to train a simple multi-layer perceptron (MLP) network to estimate the depression severity. The MLP network consists of an input layer with 250 units, a hidden layer which is a fully connected layer with 75 units, and an output layer with a single unit.

III. DATABASE

Speech data collected as part of the FORBOW (Families Overcoming Risks and Building Opportunities for Well Being) research project [27] were considered for analysis. Speech samples were collected from 526 subjects (399 mothers and 127 fathers). In these recordings, parents were asked to talk about their children for five minutes without interruption. Trained clinical assessors interviewed each participant and scored their current depression severity on the Montgomery and Asberg Depression Rating Scale (MADRS), a validated measure of depression severity [28]. The range of MADRS scores in this database is $0-21$. Further, clinical experts, blind to MADRS ratings, partitioned each 5 minute recording into 3-7 second-long segments of consistent emotion/sentiment, and provided corresponding sentiment and emotion labels for every segment. The intraclass correlation (ICC) for ratings of different clinical experts was high (ICC=0.82) showing strong agreement in the labeling. The labels considered for annotating sentiment are: Negative, Neutral and Positive. Similarly the labels considered for emotion are: Anger, Sad, Neutral and Happy. These sentiment and emotion labels were used to train the sentiment and emotion embedding based models. It is to be noted that the MADRS rating given to each 5 minute recording was assigned to all the segments obtained from that recording.

A total of 12700 segments (each of $3-7$ seconds in duration) were obtained from the 526 recordings. 11300 segments obtained from 470 recordings were considered as the train set. Remaining 1400 segments obtained from the held-out 56 recordings–i.e., all recordings corresponding to a subset of the speakers–were considered as the test set.

IV. EXPERIMENTS

A. Training Details

Note that all networks (including the embedding networks) in this paper were trained only on the training set and never on the held-out test set. All the networks were trained using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and with an initial learning rate of 0.0005. Dropout rate of 0.3 and 0.4 was considered for the convolutional and fully connected layers, respectively to avoid model overfitting. ReLU activation was used for the embedding layers. Linear activation was considered for the regression part of the output layer, and softmax activation was considered for the classification part of the output layer. All networks were trained for 50 epochs with a batch size of 32. Negative log-likelihood (NLL) and mean-squared error (MSE) loss functions were considered to train models on classification and regression tasks, respectively. For training multi-task networks, equal weightage is given to both, NLL and MSE losses.

B. Results

Table I gives root-mean-square error (RMSE) values in estimation of depression severity (MADRS scores), by considering spectral, prosodic and spectral+prosodic features to train different regression networks i.e., linear regression (Lin.-Reg.), Support vector machine based regression (SVM-Reg.), CNN based regression (CNN-Reg.), and Multi-task CNNs. It can be observed from Table I that networks trained by combining prosodic and spectral features performed better than their counterparts trained with only spectral or prosodic features. This shows that the spectral and prosodic features carry complimentary information to improve estimation of depression severity. It can also be observed that the networks trained on only regression tasks perform nearly the same (with CNNs performing slightly better than linear regression and SVM-based regression). But the multi-task training of CNN (MT-CNN) improved the performance in predicting the depression scores when compared to CNN-Reg.

The RMSE values obtained by considering sentiment and emotion embeddings for estimating depression scores are given in Table II. It can be observed from the Table II that the sentiment and emotion embeddings also carry certain depression-specific information. Further, the multi-task training of sentiment and emotion also improves the estimation of depression severity. It can also be observed from Table III that combining sentiment+emotion embeddings with the depression-specific embeddings improves the estimation of depression severity. This shows that the sentiment and emotion

TABLE II
RMSE FOR DEPRESSION SEVERITY ESTIMATION USING EMBEDDINGS FROM SENTIMENT, EMOTION, AND SENTIMENT + EMOTION (SENT. + EMO.) PREDICTION TASKS. THE EMBEDDINGS WERE LEARNED USING RAW AUDIO FEATURES LISTED.

Using Embedding	Based on	RMSE
Sentiment	Spectral	11.10
	Spectral + Prosodic	10.98
Emotion	Spectral	10.03
	Spec.+Prosodic	9.94
Sent. + Emo. (MT-CNN-SE)	Spectral	9.58
	Spectral +Prosodic	9.32

TABLE III
RMSE VALUES OF DEPRESSION SEVERITY SCORES COMBINING DEPRESSION EMBEDDINGS, AND SENTIMENT AND EMOTION EMBEDDINGS. RESULTS BASED ON MULTI-TASK MODEL. SENT., EMO., REG., CLASS. REFER TO SENTIMENT, EMOTION, REGRESSION AND CLASSIFICATION, RESPECTIVELY.

Model	**Feat**	**RMSE**
Sent. + Emo.	Spec	9.58
	Spectral + Prosodic	9.32
Reg. + Class.	Spectral	7.48
	Spectral + Prosodic	7.36
Combined	Spectral	7.01
	Spectral + Prosodic	6.93

embeddings provide important cues complementary to—and not otherwise captured by—the depression-specific embeddings. Further, the RMSE value of random classification on our test set is 12.04, which is larger compared to the RMSE values reported for any system in Tables I, II and III.

We analyzed the sentiment and emotion classification performance (in terms of accuracy (Acc. %)) obtained by considering the sentiment and emotion embeddings (results provided in Table IV). It can be observed that the multi-task training improves the sentiment and emotion classification performance. Further, considering prosodic features also helps sentiment and emotion classification based on audio signals.

V. SUMMARY

In this paper, we have proposed a multi-task learning framework, and the use of sentiment and emotion based embeddings for improving the performance of depression severity estimation from the acoustic features of short audio recordings of speech. Experimental results show that the proposed multi-task training on regression and classification tasks improves the estimation of depression severity. We also showed that a multi-task CNN trained on sentiment and emotion classification tasks attains higher sentiment and emotion classification performance compared to two individual networks. Sentiment-emotion embeddings extracted from this multi-task CNN when combined with acoustic features further improved the performance of depression severity estimation. These improvements suggest potential use of the proposed approaches in developing clinical applications.

TABLE IV
RESULTS FOR SENTIMENT (SENT.) AND EMOTION (EMO.) CLASSIFICATION (IN TERMS OF ACCURACY (ACC) %).

Model	Task	Features	Acc. (%)
Sentiment	Sent.	Spectral	39.2
		Spectral+Prosodic	40.1
Emotion	Emo.	Spectral	41.3
		Spectral+Prosodic	43.0
Sent. + Emo. (MT-CNN-SE)	Sent.	Spectral	41.8
		Spectral+Prosodic	42.6
	Emo.	Spectral	44.2
		Spectral+Prosodic	46.1

ACKNOWLEDGEMENTS

Resources used in preparing this research were provided, in part, by NSERC, the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/\#partners.

REFERENCES

[1] J. Rehm and K. D. Shield, "Global burden of disease and the impact of mental and addictive disorders," *Current psychiatry reports*, vol. 21, no. 2, pp. 10, 2019.

[2] O. J. Oluboka, M. A. Katzman, J. Habert, et al., "Functional recovery in major depressive disorder: providing early optimal treatment for the individual patient," *International Journal of Neuropsychopharmacology*, vol. 21, no. 2, pp. 128–144, 2018.

[3] World Health Organization WHO, "WHO global health days-Staying positive and preventing depression as you get older," Website, 2017.

[4] J. F. Greden, A. A. Albala, I. A. Smokler, R. Gardner, and BJ Carroll, "Speech pause time: a marker of psychomotor retardation among endogenous depressives.," *Biological Psychiatry*, 1981.

[5] Nilsonne, "Acoustic analysis of speech variables during depression and after improvement," *Acta Psychiatrica Scandinavica*, vol. 76, no. 3, pp. 235–245, 1987.

[6] HH Stassen, S Kuny, and D Hell, "The speech analysis approach to determining onset of improvement under antidepressants," *European Neuropsychopharmacology*, vol. 8, no. 4, pp. 303–310, 1998.

[7] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Interspeech*, 2012.

[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] L. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in *ICASSP*. IEEE, 2010, pp. 5154–5157.

[10] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Interspeech*, 2011.

[11] M. Valstar, B. Schuller, K. Smith, F. Eyben, et al., "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proc. ACM workshop on Audio/visual emotion challenge*, 2013, pp. 3–10.

[12] 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 *Proc. workshop on audio/visual emotion challenge*, 2014, pp. 3–10.

[13] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, et al., "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proc. Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACM conference on Multimedia*, 2010, pp. 1459–1462.

[15] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of biomedical informatics*, vol. 83, pp. 103–111, 2018.

[16] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, 2019.
[17] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," *Proc. ICML*, 1993.
[18] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.
[19] Md Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *Proc. NAACL-HLT*, 2019, pp. 370–379.
[20] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L. P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. ACL*, 2017, pp. 873–883.
[21] I. Sheikh, S. H. Dumpala, R. Chakraborty, and S. K. Kopparapu, "Sentiment analysis using imperfect views from spoken language and acoustic modalities," in *Proc. Grand Challenge and Workshop on Human Multimodal Language*, 2018, pp. 35–39.
[22] H. Jiang, B. Hu, Z. Liu, L. Yan, T. Wang, F. Liu, H. Kang, and X. Li, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Communication*, vol. 90, pp. 39–46, 2017.
[23] X. Cheng, X. Wang, T. Ouyang, and Z. Feng, "Advances in emotion recognition: Link to depressive disorder," in *Mental Disorders*. IntechOpen, 2020.
[24] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3d facial expressions," *arXiv preprint arXiv:1811.08592*, 2018.
[25] A. U. Hassan, J. Hussain, M. Hussain, M Sadiq, and S. Lee, "Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression," in *Proc. International Conference on Information and Communication Technology Convergence*. IEEE, 2017, pp. 138–140.
[26] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. 99, no. 7, 2016.
[27] R. Uher, J. Cumby, L. E. MacKenzie, J. Morash-Conway, J. M. Glover, A. Aylott, L. Propper, S. Abidi, A. Bagnell, B. Pavlova, et al., "A familial risk enriched cohort as a platform for testing early interventions to prevent severe mental illness," *BMC psychiatry*, vol. 14, no. 1, pp. 344, 2014.
[28] S. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change.," *The British Journal of Psychiatry*, 1979.