Deep-learning multimodal system for anxiety and depression detection

Brian Diep[1, 2, *], Marija Stanojevic[1], Jekaterina Novikova[1]

[1] Winterlight Labs, Toronto, ON, Canada
[2]Faculty of Arts and Science, University of Toronto, Toronto, ON, Canada
*Presenting Author; Disclosures: All authors are/were employees of Winterlight Labs, Inc.

Background

Depression and anxiety are two of the most common psychiatric disorders globally. [1]. Although these disorders can negatively impact the quality of life for patients, early diagnosis and intervention can significantly benefit outcomes for the patient [2. 3]. Unfortunately. current approaches to diagnosing depression or anxiety through clinical assessment can pose a high-burden to patients seeking care and can suffer from issues with subjectivity [4]. Speech-based biomarkers provide an interesting avenue for the future of diagnosis or monitoring of depression and anxiety. Patterns in the acoustic and linguistic content of speech can be used to differentiate depressed/anxious and non-depressed/anxious individuals [5].

Study Objective:

Present a framework for using speech as a biomarker in the diagnosis of depression and anxiety

Methods

3543 participants with 4209 unique samples were recruited over Amazon Mechanical Turk (mTurk) completed speech task assessments. These participants were prompted to complete one-minute tasks where they were asked to describe events or experiences. Audio recordings were transcribed and analyzed using signal and natural language processing (NLP) to derive >500 acoustic and linguistic measures. The participant's depression and anxiety symptoms were also collected following the Patient Health Questionnaire-8 and Generalized Anxiety Disorder-7 scales. Scale scores were converted to "soft" binary labels for depression and anxiety diagnosis by taking a score of 10 as a cutoff. Our classifier model is then trained on this data to do binary classification of the diagnosis labels. To evaluate our model, we used 5-fold cross validation where precision, recall, and F1-score for each diagnosis label were measured and averaged over each validation fold..

Model Architecture

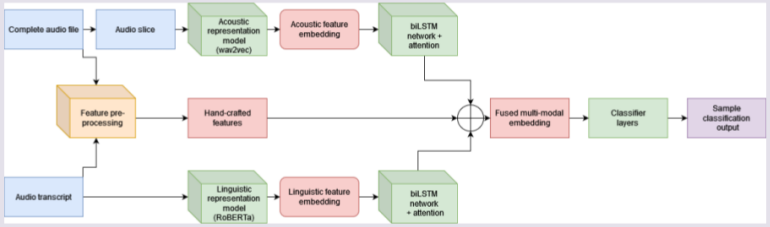

Figure 1 . Model architecture diagram of classification model, adapted from [6]

This architecture involves the multi-modal fusion of the acoustic and linguistic dimensions of speech. We have two parallel branches that create deep-learned representations of speech using Wav2vec 2.0 and RoBERTa.
These representations along with our hand-crafted features form a comprehensive representation of speech which is used to predict the binary classification label of any input speech sample.

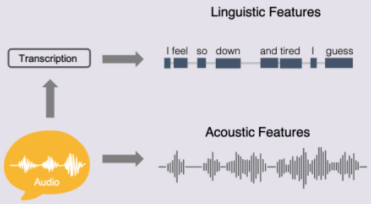

Figure 2. Schematic representation of how acoustic and linguistic speech features are extracted from raw audio.

Results

Table 1. Anxiety and depression classification results

	Anxiety						Depression					
	Hand-crafted only			Hand-crafted + deep-learned			Hand-crafted only			Hand-crafted + deep-learned		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
< 10	0.81	0.65	0.72	0.76	0.72	**0.73**	0.73	0.78	0.75	0.77	0.83	**0.80**
≥ 10	0.28	0.41	0.33	0.37	0.42	**0.40**	0.31	0.42	0.35	0.48	0.39	**0.43**
All			0.54			**0.57**			0.58			**0.63**

Discussion

The results of our experiments are displayed here. We find that augmenting hand-crafted features with deep-learned features improves our overall classification F1 score compared to a baseline of hand-crafted features alone from 0.58 to 0.63 for depression and from 0.54 to 0.57 for anxiety. The results show that the inclusion of deep-learned features enriches the representation by adding properties that are not fully captured by the hand-crafted features, improving the detection of depression and anxiety.
Classification of depression was better than anxiety which could potentially be because of data imbalance. As only had 12.8% of GAD-7 scores as opposed to 25.3% of PHQ-8 scores in our dataset had scores above the soft diagnosis threshold. Future studies with an expanded dataset and more samples from higher severity-bands of depression and anxiety can help validate this approach.

Key Findings:

1. Speech is a rich source of information that can be used for the diagnosis of depression and anxiety

2. Crowd-sourcing is a valuable way of gathering samples for biomarker development

3. Combining deep-learned and hand-crafted speech features can improve the classification of depression and anxiety

References

[1] Villarroel MA, Terlizzi EP. Symptoms of Depression Among Adults: United States, 2019. *NCHS Data Brief*, No. 379 (2020).
[2] Dadds, M. R., Spence, S. H., Holland, D. E., Barrett, P. M., & Laurens, K. R. (1997). Prevention and early intervention for anxiety disorders: a controlled trial. *Journal of Consulting and Clinical Psychology*, *65*(4), 627.
[3] Reynolds III, C. F., Cuijpers, P., Patel, V., Cohen, A., Dias, A., Chowdhary, N., ... & Albert, S. M. (2012). Early intervention to reduce the global health and economic burden of major depression in older adults. *Annual review of public health*, *33*, 123-135.
[4] Nease, D. E., & Malouin, J. M. (2003). Depression screening: a practical strategy. *Journal of Family Practice*, *52*(2), 118-126.
[5] Pope, B., Blass, T., Siegman, A. W., & Raher, J. (1970). Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology*, *35*(1p1), 128.
[6] Toto, E., Tlachac, M. L., & Rundensteiner, E. A. (2021, October). Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 4145-4154).

Presented at A Participatory Approach to AI for Mental Health (PAI4MH), 2022