Detecting Anxiety and Depression from Phone Conversations using x-vectors

Namhee Kwon[1]*, Shahruk Hossain*[1]*, Nate Blaylock*[1]*, Henry O'Connell*[1]*,*
Naomi Hachen[2]*, Joseph Gwin*[2]

[1]Canary Speech, USA
[2]Best Buy Health, USA

{namhee, shahruk, nate}@canaryspeech.com, {naomi.hachen, joseph.gwin}@bestbuy.com

Abstract

We developed a model for detecting anxiety and depression from telephony recordings between a customer and a representative at a call center using vocal features and a deep neural network. Our binary classification model using x-vectors outperformed the use of the other acoustic features such as i-vectors and openSMILE features, as well as linguistic or text-based features. Our models were built based on self-reported scores: GAD-7 for anxiety and PHQ-8 for depression. Especially, the anxiety model's performance is very similar to the GAD-7 score's screening accuracy. A prior study compared self-reported GAD-7 scores to an actual mental health professional's diagnosis of anxiety disorder and reported sensitivity and specificity of 0.74 and 0.54 respectively, and our model showed a sensitivity of 0.70 and a specificity of 0.54. This study exhibits the potential of voice analysis on topic-independent speech, particularly from 8 kHz phone conversations, to identify anxiety and depression.

Index Terms: speech analysis, anxiety detection, depression detection, mental health, x-vectors, telephony

1. Introduction

Mental health evaluation is a vital topic of social interest and concern. This has motivated research into the automatic identification of mental health issues such as depression and anxiety, using different modalities, such as audio and video features [1, 2]. Some of these studies used specially designed questions or prompts [3, 4], while others relied on open conversations or spontaneous speech [5, 2]. The audio used is often from a "mental health evaluation interview" setting and recordings are wide-band (16 kHz or more).

In this study, we developed a neural network model that used phone conversations, recorded at 8 kHz, for evaluating a caller's anxiety and depression. This was motivated by the fact such a model could be applied in call centers and automated health check-up calls, especially for older adults. Since telephony audio is sampled at 8 KHz, the maximum frequency bin resolvable (Nyquist Frequency) is 4 kHz; higher sampling rates may have offered better audio fidelity and features for frequencies higher than 4KHz, but we were constrained by the application environment (i.e. telephone interviews). While some prior studies have analyzed the *state* anxiety such as anxiety during public speech [6, 7], we focused on *trait* anxiety which is a measure of the stable tendency of the anxiety in a person, separate from a momentary situation one might be in [8]. Trait anxiety may manifest more pervasively in speech compared to momentary or state anxiety - which as the name suggests - may be observable in only part of the telephone conversation.

We explored several features for this task including i-vectors [9] and x-vectors [10] which were originally developed for speaker identification and diarization. These vectors or embeddings encode several aspects of a speaker's voice [11], and have been used in applications besides speaker identification including language identification [12], detecting early symptoms of Parkinson's and Alzheimer's [13, 14, 15] and sentiment evaluation [16, 17]. We utilized publicly available i-vector and x-vector models from the Kaldi website [18] for our experiments. We also evaluated the use of linguistic features (speech rate, articulation rate, pause rate etc.) and features from the openSMILE toolkit [19, 20] for comparison.

2. Data

We partnered with the call-center of a large telehealth organization and with the consent of customers calling in, collected conversations between the customer (caller) and a representative (agent), from November 2021 through January 2022. The customers primarily called in regarding account billing issues. Callers were randomly selected at the end of a call to participate in the program and were offered a small bill credit for volunteering. Callers were told that they could drop out at any time and still receive the compensation.

2.1. Labels and Annotations

We used the self-reported score on the GAD-7 questionnaire for anxiety and PHQ-8 questionnaire for depression as the gold-standard label. The GAD-7 (General Anxiety Disorder-7) is a standard self-reported scale for screening anxiety disorder proposed by Spitzer et al. in 2006 [21]. It is composed of 7 questions asking how often the person has been bothered by specific feelings over the past 2 weeks. The total score, y, ranges between 0 to 21, calculated by summing each question's answer which are on a 4-point scale (0-3). Spitzer suggested cutoff scores for 4 categories to interpret the GAD-7 scores: minimal ($y < 5$), mild ($y < 10$), moderate ($y < 15$), and severe ($y \geq 15$). A study by Beard et al. in 2014 [22] compared the efficacy of the GAD-7 for anxiety disorder detection against a mental health professional's diagnosis and reported sensitivity and specificity of 0.74 and 0.54 respectively. Sensitivity and Specificity are commonly used in the medical field to evaluate diagnostic efficacy. They are defined as the proportion of correctly diagnosed positive and negative labels respectively; that is sensitivity is the recall of positive labels and specificity is recall of the negative labels and have values between 0 and 1. Note that the comparison between the self-reported GAD-7 scores and the diagnoses of a mental health professional was done on binary categories: anxiety disorder when $y \geq 10$, and no or minimal anxiety disorder when $y < 10$, not on 4 categories or continuous scores.

The PHQ-8 is an 8-item questionnaire on depressive symptoms. It is based on the PHQ-9 (Patient Health Questionnaire-

9) by excluding its final question, which asks about suicidal thoughts. The PHQ-9 [23] is a standard screening questionnaire, with answers on a 4-point scale for each item, totaling between 0 and 27 (24 for PHQ-8). Previous research [24] showed a very strong correlation of 0.996 between PHQ-8 and PHQ-9. Both PHQ-8 and PHQ-9 suggest a cut-off score of 10 for screening major depressive disorder. PHQ-9 is reported to have a sensitivity of 0.74 and a specificity of 0.91 using the suggested cut-off score when compared to a health professional's diagnosis of depression disorder [25].

At the end of the call, the representatives asked GAD-7 and PHQ-8 questions to the callers orally. GAD-7 and PHQ-8 are proposed as a self-administered tool but are also used in in-person interviews or telephone assessments. A strong concordance between telephone- and self-administered assessments was reported in [26] and the evidence of equivalence between telephone and in-person interviews was reported in [27].

Recordings having responses to all the questions in the GAD-7 and PHQ-8 questionnaire were double-annotated by native English speakers after being familiarized with a standard annotation manual. Any discrepancies were discussed and adjudicated to finalize the label. The inter-annotator agreement was measured with Cohen's kappa statistic at 0.950, which represents a very reliable annotation.

2.2. Data Preprocessing

We de-identified all recorded calls and discarded the channel containing the agent's audio. Any personal identifiable information was redacted and large segments (or speech turns) with identifiable information were removed from the dataset. The part of the call where the GAD-7 and PHQ-8 questionnaires were administered was also removed; so we only used the first part of the call where callers discussed the billing issues for training and evaluating our models. Any calls having less than 40 seconds of audio (after the aforementioned filtering) were also discarded. The mean audio recording time was 124 seconds. The total number of audio calls after all filtering steps was 291 for anxiety and 295 for depression labels.

The detailed label distribution by applying the suggested cut-off scores is presented in Table 1. The moderate or severe label categories were under-represented in the distribution for both GAD-7 and PHQ-8, which is consistent with the general population of anxiety and depression. The GAD-7 label and the PHQ-8 label were very positively correlated; the Pearson's correlation coefficient is 0.74 ($p <$ 1e-10).

Table 1: *Label Distribution*

GAD-7 Label	Score	Count	Percent
Minimal	$0 \leq y < 5$	173	60%
Mild	$5 \leq y < 10$	68	23%
Moderate	$10 \leq y < 15$	33	11%
Severe	$y \geq 15$	17	6%
Total		291	100%

(a) *Anxiety Label*

PHQ-8 Label	Score	Count	Percent
Minimal	$0 \leq y < 5$	192	61%
Mild	$5 \leq y < 10$	66	20%
Moderate	$10 \leq y < 15$	24	9%
Mod. Severe	$15 \leq y < 20$	10	3%
Severe	$y \geq 20$	3	1%
Total		295	100%

(b) *Depression Label*

3. Approach

3.1. Proposed Model

Our proposed model architecture is presented in Figure 1. The model consists of a pretrained x-vector [10] extractor and a trainable classifier stacked on top. The model analyzes fixed length segments of audio at a time and predicts a binary label for each. The segment length was heuristically fixed at 36 seconds after preliminary experiments and observing validation loss using segment lengths between 6 and 60 seconds. The binary label corresponds to the GAD-7 or PHQ-8 score with a cutoff of 10 as described in Section 2.1, i.e. 1 if $y \geq 10$, 0 otherwise. The predicted labels for all analyzed segments for a given speaker are combined through majority voting to produce the final label.

We used the freely available *SRE16 Xvector Model* on the Kaldi website, which was trained on 8 kHz audio from several datasets including *Switchboard* [28], *Mixer 6* [29] and *NIST SREs* [30]. The input to the x-vector model are feature frames containing 23 Mel Frequency Cepstral Coefficients (MFCCs) extracted using a frame length and shift of 25 and 10 milliseconds respectively. Additionally, the MFCC frames are normalized using Cepstral Mean Variance Normalization (CMVN) before they are fed to the network. The x-vector model itself is a neural network composed of 5 initial 1-D convolutional layers (*TDNN layers* in Kaldi), a stats pooling layer, and a final fully connected layer. All convolutions have symmetric left-right contexts. The convolution layers have kernels of size 5, 3, 3, 1, and 1 respectively. The second and third convolutions have dilation rates of 3 and 2 respectively. All convolutions are followed by ReLU activation and batch normalization [31]. The first four convolution layers have 512 filters, while the fifth has 1500. The stats pooling layer computes the mean and standard deviation of each feature dimension in the fifth convolutional layer's output, producing a feature vector with 3000 dimensions. The stats are computed using a sliding window of size and stride equal to 150 acoustic feature frames, or 1.5 seconds of audio. This is the default setting used while extracting x-vectors in Kaldi. The feature vector composed of the means and variances is transformed by a final fully connected layer to produce the final emitted 512 dimensional x-vectors.

Before being fed to the classifier, every four x-vectors are averaged to give a more robust x-vector that spans 6 seconds of audio. The classifier architecture is a convolutional network with three 1-D convolutions, all with a kernel size of 3 and 128, 128 and 64 filters respectively. This is followed by a fully connected layer with 32 nodes, and an output layer with 2 nodes and softmax activation. As is common practice, each convolution is followed by batch ReLU activation, batch normalization, and dropout. The entire model (x-vector + classifier) has 4.5 million parameters, of which 0.28 million are trainable (classifier).

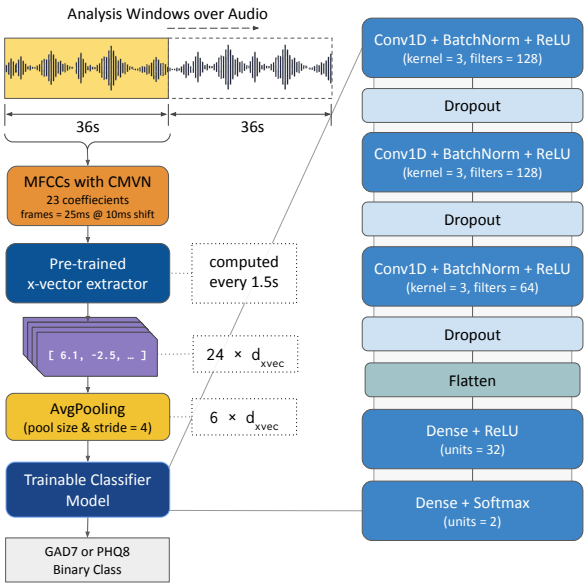

Figure 1: *Classifier Architecture using x-vectors*

3.2. Other Models

We explored text-based and linguistic feature models based on the transcripts from automatic speech recognition (ASR) system. NLP approaches have been applied to mental illness detection [32] but the ASR accuracy affects the model performance when they are applied to spoken language data.

We transcribed all calls using Amazon's Transcribe ASR service and extracted "language" features identifying various characteristics of lexical, syntactic and semantic content of the speech. This included lexical complexity defined by n-gram probability (using a language model generated using the Librispeech dataset [33]), word usage score (SMOG reliability score) [34], word ambiguity and familiarity score based on linguistic research [35]. The semantic features included subjective word score and depression cue word score generated by referencing hand-generated dictionaries. Linguistic features included speech rate, vowel length, voice/pause ratio, filler word ratio, and repetitive word ratio. The final vector size of the "language" features was 53.

We also evaluated the use of word and sentence embeddings generated by a pretrained BERT model [36]. BERT is a context-aware language representation pretrained from unlabeled text. We generated an embedding per call first by returning the global attention output ([CLS] token embedding) and second by averaging all words' embeddings. Using the BERT base model whose embedding size is 768, we used the final layer of CLS embedding for the global vector model (BERT-CLS), so the final feature size was 768. However, for the averaging model, we used the final 4 layers of word vectors for averaging so the output vector size was 3072 (4 * 768) (BERT-AVG-4). We limited the max input sequence length at 128 words

In addition to the ASR-based language and BERT features, we experimented acoustic and prosodic features using the openSMILE toolkit (widely used in emotion recognition) [37]. With different configurations in openSMILE, various feature sets were generated and tested. They were generated per call (per assessment) as in ASR-based features. Specifically, we used the eGeMaps [37], ComParE [38] and IS09 [39] feature sets from openSMILE; they comprise of several different low level audio descriptors and higher level statistics over them (mean, standard deviation, kurtosis, delta etc.)

We also experimented with i-vectors [9]. Because we did not get a publicly available i-vector model trained on 8kHz data, we used the 16kHz pre-trained model from the Kaldi website trained on VoxCeleb [40]. We upsampled the 8kHz input into 16kHz and generated the i-vector representation.

We applied a fully connected deep neural network architecture for the above features. The number of hidden nodes, the layer depth, and the dropout rate are tuned by the balanced accuracy of the development set using hyperband search [41]. We used the ReLU activation and added a sigmoid layer at the end.

4. Experiments

4.1. Set Up

We trained two separate models for anxiety and depression and performed 10-fold cross validation for each label. We split the data into 10 stratified folds, with the same label distribution as the full dataset. For each training run, we selected one fold as a dev set for auto-tuning hyper-parameters and model selection. We chose another as the test set to evaluate the best dev set model. The other folds were used as the train set. We ran training 10 times so that each fold was chosen as the test set once. We reported the performance on the combined test result from the 10 experiments in Section 4.3.

4.2. Training

All experiments were performed using the Tensorflow framework [42]. The Kaldi x-vector model was converted for use in Tensorflow using the kaldi-tflite codebase [43]. For each mini-batch, the training pipeline randomly selected a segment of audio from the shuffled set of speakers. The appropriate features were generated on-the-fly for these segments and provided to the models for training.

For experiments with openSMILE features and other text-based features described in Section 3.2, we generated a fixed length of feature vector per call and fed it into a fully connected (dense) model.

For all experiments, we used the Adam optimizer algorithm [44] with an initial learning rate of 10^{-4} and a cosine decay schedule (with restarts). We used binary focal cross-entropy loss which was reported to be more effective for datasets with label imbalance [45].

4.3. Results

The experimental results are presented in Table 2. We measure the sensitivity and specificity to directly compare with the GAD-7 and PHQ-8's reliability. There is a trade-off between sensitivity and specificity and each model shows a different preference for sensitivity over specificity. The performance comparison between models is mainly done by the balanced accuracy (a.k.a. unweighted accuracy) which is equal to the average recall of all classes. The balanced accuracy is a measure often used for data with an imbalanced label distribution and the by-chance model's balanced accuracy is 0.5.

The anxiety and depression model show very similar sensitivity and specificity and our proposed models using x-vector outperform other baseline models using openSMILE or text-based features for both labels. Considering that our final goal is

to make a screening tool for anxiety and depression where sensitivity is more important than specificity, it is encouraging that the x-vector model shows higher sensitivity than other models.

Table 2: *Classifier Model Performance*

Features	Model	BalAcc	Sensitivity	Specificity
x-vector	**CNN**	**0.62**	**0.70**	**0.54**
Language	Dense	0.60	0.62	0.58
BERT-CLS	Dense	0.50	0.54	0.47
BERT-Avg-4	Dense	0.59	0.60	0.59
i-vector	Dense	0.45	0.21	0.70
opensmile (eGeMaps)	Dense	0.45	0.40	0.50
opensmile (ComParE)	Dense	0.44	0.44	0.44
opensmile (IS09)	Dense	0.44	0.30	0.59

(a) *Anxiety Model Comparison*

Features	Model	BalAcc	Sensitivity	Specificity
x-vector	**CNN**	**0.62**	**0.73**	**0.50**
Language	Dense	0.56	0.54	0.58
BERT-CLS	Dense	0.60	0.63	0.58
BERT-Avg-4	Dense	0.48	0.52	0.44
i-vector	Dense	0.57	0.39	0.74
opensmile (eGeMaps)	Dense	0.54	0.59	0.49
opensmile (ComParE)	Dense	0.51	0.33	0.70
opensmile (IS09)	Dense	0.48	0.28	0.69

(b) *Depression Model Comparison*

5. Conclusions

We developed an anxiety detection model and a depression detection model. We trained our model using x-vectors extracted from 8 kHz audio of phone conversations between a customer and a call center representative. We compared the use of x-vectors with other features including i-vectors, linguistic features such as speech rate, articulation rate, and other text-based features such as lexical complexity and embeddings from BERT. We used data collected from a call-center that were labelled using the scores from the GAD-7 and PHQ-8 questionnaires taken by the customers. The model using x-vectors outperformed the rest, showing sensitivity of 0.70 and a specificity of 0.54 for anxiety and sensitivity of 0.73 and specificity of 0.50 for depression. The anxiety model performance is very similar to the reported performance of the GAD-7 questionnaire itself when compared to a mental health professional's diagnosis (0.74 and 0.54 for sensitivity and specificity respectively). The depression model performance is also close to the PHQ-8 questionnaire's performance in sensitivity (0.73 vs. 0.74) although the specificity is lower (0.50 vs. 0.91). This shows that this voice model can be developed as a anxiety and depression screening tool over phone conversations. We plan to perform more detailed feature analysis, combine x-vectors with other features and add more data to improve the performance of the model.

6. References

[1] *AVEC '16: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: Association for Computing Machinery, 2016.

[2] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, and J. A. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 2, pp. 1–26, 2018.

[3] S. Kim, N. Kwon, H. O'Connell, N. Fisk, S. Ferguson, and M. Bartlett, ""how are you?" estimation of anxiety, sleep quality, and mood using computational voice analysis," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2020, Montreal, QC, Canada, July 20-24, 2020*. IEEE, 2020, pp. 5369–5373.

[4] 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*, 2016, pp. 35–42.

[5] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker *et al.*, "From joyous to clinically depressed: Mood detection using spontaneous speech." in *FLAIRS Conference*, vol. 19. Citeseer, 2012.

[6] E. H. Nirjhar, A. Behzadan, and T. Chaspari, "Exploring bio-behavioral signal trajectories of state anxiety during public speaking," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1294–1298.

[7] M. N. Sakib, E. H. Nirjhar, K. Feng, A. Behzadan, T. Chaspari, and T. Chaspari, "Exploring individual differences of public speaking anxiety in real-life and virtual presentations," pp. 1–1.

[8] N. S. Endler and N. L. Kocovski, "State and trait anxiety revisited," *Journal of anxiety disorders*, vol. 15, no. 3, pp. 231–245, 2001.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[11] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 726–733.

[12] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, 2018, pp. 105–111.

[13] L. Jeancolas, D. Petrovska-Delacrétaz, G. Mangone, B.-E. Benkelfat, J.-C. Corvol, M. Vidailhet, S. Lehéricy, and H. Benali, "X-vectors: New quantitative biomarkers for early parkinson's disease detection from speech," *Frontiers in Neuroinformatics*, vol. 15, p. 578369, 2021.

[14] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[15] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, "Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity." in *INTERSPEECH*, 2020, pp. 2177–2181.

[16] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[17] L. F. Parra-Gallego and J. R. Orozco-Arroyave, "Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments," *Digital Signal Processing*, vol. 120, p. 103286, 2022.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[20] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224

[21] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Arch Intern Med.*, vol. 166, no. 10, pp. 1092–1097, 2006.

[22] C. Beard and T. Björgvinsson, "Beyond generalized anxiety disorder: Psychometric properties of the gad-7 in a heterogeneous psychiatric sample," *Journal of Anxiety Disorders*, vol. 28, no. 6, pp. 547–552, 2014.

[23] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The phq-9: validity of a brief depression severity measure," vol. 16, no. 9, pp. 606–613. [Online]. Available: https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[24] Y. Wu, B. Levis, K. E. Riehm, and et al., "Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis," vol. 50, no. 8, pp. 1368–1380, edition: 2019/07/12 Publisher: Cambridge University Press.

[25] B. Arroll, F. Goodyear-Smith, S. Crengle, J. Gunn, N. Kerse, T. Fishman, K. Falloon, and S. Hatcher, "Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population," vol. 8, no. 4, p. 348. [Online]. Available: http://www.annfammed.org/content/8/4/348.abstract

[26] A. Pinto-Meza, A. Serrano-Blanco, M. T. Peñarrubia, E. Blanco, and J. M. Haro, "Assessing depression in primary care with the phq-9: can it be carried out over the telephone?" *Journal of General Internal Medicine*, vol. 20, no. 8, pp. 738–742, Aug 2005.

[27] Y. Conwell, A. Simning, N. Driffill, Y. Xia, X. Tu, S. P. Messing, and D. Oslin, "Validation of telephone-based behavioral assessments in aging services clients," *International Psychogeriatrics*, vol. 30, no. 1, pp. 95–102, Jan 2018.

[28] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[29] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 2010.

[30] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The nist speaker recognition evaluation–overview, methodology, systems, results, perspective," *Speech communication*, vol. 31, no. 2-3, pp. 225–254, 2000.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[32] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," vol. 5, no. 1, p. 46. [Online]. Available: https://doi.org/10.1038/s41746-022-00589-7

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[34] K. L. Grabeel, J. Russomanno, S. Oelschlegel, E. Tester, and R. E. Heidel, "Computerized versus hand-scored health literacy tools: a comparison of simple measure of gobbledygook (smog) and flesch-kincaid in printed patient education materials," *Journal of the Medical Library Association: JMLA*, vol. 106, no. 1, p. 38, 2018.

[35] S. Cho, N. Nevler, S. Ash, S. Shellikeri, D. J. Irwin, L. Massimo, K. Rascovsky, C. Olm, M. Grossman, and M. Liberman, "Automated analysis of lexical features in frontotemporal degeneration," *Cortex*, vol. 137, pp. 215–231, 2021.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[37] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, 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, pp. 1–1, 01 2015.

[38] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, 2016, pp. 2001–2005.

[39] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," 2009.

[40] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[41] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6765–6816, 2017.

[42] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[43] S. Hossain, "Kaldi models in tensorflow lite," https://github.com/shahruk10/kaldi-tflite, 2021.

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

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017. [Online]. Available: https://arxiv.org/abs/1708.02002