INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Domain Adaptation for Enhancing Speech-based Depression Detection in Natural Environmental Conditions Using Dilated CNNs

Zhaocheng Huang[1], *Julien Epps*[1], *Dale Joachim*[2], *Brian Stasak*[1,2], *James R. Williamson*[3], *Thomas F. Quatieri*[3]

[1]School of Electrical Engineering and Telecommunications, UNSW Sydney, Australia
[2]Sonde Health, Boston MA, USA
[3]Massachusetts Institute of Technology Lincoln Laboratory, Lexington MA, USA

zhaocheng.huang@unsw.edu.au, j.epps@unsw.edu.au, djoachim@sondehealth.com, b.stasak@unsw.edu.au, jrw@ll.mit.edu, quatieri@ll.mit.edu

Abstract

Depression disorders are a major growing concern worldwide, especially given the unmet need for widely deployable depression screening for use in real-world environments. Speech-based depression screening technologies have shown promising results, but primarily in systems that are trained using laboratory-based recorded speech. They do not generalize well on data from more naturalistic settings. This paper addresses the generalizability issue by proposing multiple adaptation strategies that update pre-trained models based on a dilated convolutional neural network (CNN) framework, which improve depression detection performance in both clean and naturalistic environments. Experimental results on two depression corpora show that feature representations in CNN layers need to be adapted to accommodate environmental changes, and that increases in data quantity and quality are helpful for pre-training models for adaptation. The cross-corpus adapted systems produce relative improvements of 29.4% and 17.2% in unweighted average recall over non-adapted systems for both clean and naturalistic corpora, respectively.

Index Terms: Depression detection, deep learning, domain adaption, environmental noise, mental health, smart devices.

1. Introduction

Depression is a common and costly condition, affecting 10%-15% of the global population [1]. To help ease this serious health concern, an objective, passive, ubiquitous, convenient, and cost-effective device for capturing cognitive-behavioral information would be a compelling tool for research and clinical practice [2–4]. Currently, over 80% of US adults own smart devices (e.g. phone, tablet, watch) [5], speech signals from which could be used for depression screening. This provides an unprecedented opportunity to expand access to much needed medical help for depressed individuals. Although research to date has shown considerable potential [5–8], this area remains challenging and relatively understudied.

Speech production involves complex cognitive planning and motoric actions that are often impeded by depression in a variety of ways [4], e.g. muscle tension disturbances and cognitive impairments [9], causing articulatory incoordination [10] and abnormal phoneme rates [11]. Accordingly, a few effective frameworks have been proposed to exploit speech articulation-based information for depression detection, such as vowel space area [12], speech landmarks [13, 14], vocal tract coordination (VTC) features [10, 15], and FVTC-CNN (Full VTC-Convolutional Neural Networks) [16].

The effectiveness of features generated from these frameworks can be greatly reduced when applied to cross-corpus data. Often a particular feature set, or model, that performs well on one dataset may generalize poorly to others, especially if there is a mismatch in how the data were generated. This happens, for example, when one data set is collected in real life environments and another is collected in a controlled laboratory setting. This challenge of cross-corpus evaluation is a common one in affective computing, and often results in close to chance-level performance on mismatched datasets [17–20].

In this paper, we investigate domain adaptation based on a deep learning framework to bridge the gap for cross-corpus experiments. We also compare adaptation from a large dataset with more variable ground truth quality with adaptation from a small dataset containing high quality ground truth.

2. Related Work

Automatic assessment of depression from human voice has gained increasing interest recently [21–23]. Many systems have been proposed, but there has been a recent shift towards deep learning approaches due to strong depression classification /prediction results [24–25]. To date, the majority of studies have utilized clean speech recordings collected in controlled laboratory settings. By contrast, depression screening 'in-the-wild' (i.e. on smartphones in naturalistic environments) remains challenging and relatively less explored [28]. Consequently, depression screening systems built using clean speech data are less likely to generalize well 'in-the-wild'. This weakness in generalization is due to a wide spectrum of variability in real life data collections (e.g. demographics, speech tasks, recording devices/environments, and annotation standards).

One feasible solution to the cross-corpus discrepancy is domain adaptation. The idea originates from transfer learning [27], and is not uncommon in speech-related tasks such as universal background models in speaker recognition [28]. Recently, domain adaptation based on deep learning has attracted increasing attention [29–30], because deep learning is effective in learning useful feature representations that are transferrable across different tasks, e.g. SoundNet in speech, AlexNet in image, and BERT in text [29–33].

There have been a few studies investigating generalization and transferability for domain adaptation [34–38]. For instance,

[1]Distribution Statement A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the cooperative research & development agreement under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the cooperative research & development agreement.

http://dx.doi.org/10.21437/Interspeech.2020-3135

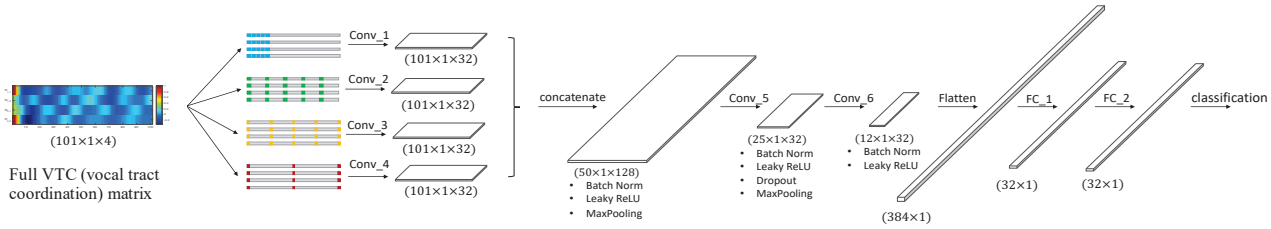

Figure 1: *The FVTC-CNN structure for exploiting vocal tract coordination, reproduced from* [16]. *Conv-1 to Conv-4 capture vocal tract coordination (represented by delayed correlations of feature contours, e.g. formant contours across time) at different time scales, Conv-5 and Conv-6 learns more abstract depression-specific information, and FC-1 and FC-2 perform classification modelling. Conv-1 to Conv-6 and FC-1, 2 can be adapted separately or collectively. The numbers of filters or neurons at each layer are annotated.*

it was found that deeper layers in CNNs can capture more task-specific information [33], and have increased linear separability than preceding layers [36]. However, the way in which transferrable information is learned by deep learning systems remains relatively unexplored in speech-related applications, including depression detection. Another under-studied area for adaptation is the trade-off between *quantity* and *quality* of data for pre-training. Quantity-quality trade-offs are to some extent inevitable in depression detection, where high quality clinically validated data are usually small in size and an ever-increasing amount of data can be collected on smartphones. The latter data often contains noisy speech data and poor quality labels.

3. Methods

3.1. System Overview

This study investigates domain adaptation for depression detection using the FVTC-CNN (full vocal tract coordination – convolutional neural networks) framework [16]. FVTC-CNN consists of two parts, i.e. an image-like FVTC matrix (Fig. 1) and a dilated CNN. The FTVC matrix consists of delayed auto- and cross-correlations from feature contours over time (e.g. formants). Those correlations were found to be associated with motor incoordination of vocal tract activities [10]. As shown in Fig. 1, a dilated CNN is used to learn the image-like full VTC matrix, which is calculated per audio file. This framework is used for three reasons: (1) it captures speech motor coordination information that could be more robust to various noise and handset variability, and hence more performant in naturalistic environments [16]; (2) the deep learning framework has higher feasibility and potential for domain adaptation than non-deep learning frameworks; and (3) the framework contains a straightforward structure, which allows interpretability for cross-corpus experiments.

Consider two datasets: D_1 and D_2, each divided into train, development and test partitions, i.e. D_i^{train}, D_i^{dev}, and D_i^{test}, where $D_i \in \{D_1, D_2\}$. The domain adaptation follows two steps:

i. A pre-trained model $\mathcal{M}_i$ is built from one corpus D_i by training on D_i^{train} and then optimizing on D_i^{dev}.

ii. Weights in the pre-trained model $\mathcal{M}_i$ are then updated on the second corpus D_j by adapting on the training partition D_j^{train}, leading to $\mathcal{M}_{i\to j}$. The frozen and adapted layers in CNN were controlled by the proposed adaptation strategies. The updated model $\mathcal{M}_{i\to j}$ is optimized on D_j^{dev}, and tested on the test partition D_j^{test}.

This adaptation approach is beneficial, because the resultant model $\mathcal{M}_{i\to j}$ benefits from additional training data, while also containing information from different datasets, which aids in mismatch compensation. For instance, $\mathcal{M}_{i\to j}$ can be knowledgeable about both clean high-quality depressed speech and noisy speech data collected from various devices in naturalistic environments. Also, it is observed that $\mathcal{M}_i$ offers a good starting point for training a model on other datasets. This is important as depression corpora are often unbalanced and relatively small in size, leading to a higher risk of achieving local minima due to poor initializations.

3.2. Proposed Adaptation Strategies

In the dilated CNN framework shown in Fig. 1, it is reasonable to expect that not all the layers are suitable for adaptation, since some layers may carry similar depression-related information across corpora. Therefore, it is beneficial for these layers to remain frozen during adaptation. However, how best to perform domain adaptation remains unclear. In this study, we propose three adaptation strategies, namely layer-wise adaptation and two types of accumulative adaptation, FirstN and LastN.

- *Layer-wise* adaptation: a single layer with trainable weights is updated.
- *FirstN* adaptation: the first *N* layers are jointly adapted.
- *LastN* adaptation: the last *N* layers are jointly adapted.

In FVTC-CNN, there are eight trainable layers of interest, i.e., six convolutional layers (i.e., Conv-1, Conv-2, …, Conv-6) and two fully connected layers (i.e., FC-1 and FC-2). The batch normalization layers are kept frozen during adaptation based on preliminary findings of no gain, as in [37]. The 'Conv' layers learn feature representations, whereas 'FC' layers learn classification models. This raises an interesting question regarding the CNN substructure: should we update the weights for only the feature representation or for only the classifier models? This question will be answered by investigating the proposed adaptation strategies.

Layer-wise adaptation examines the effectiveness of adapting particular informative layers that mitigate mismatches between different corpora. *FirstN* adaptation examines the cumulative effect of levels of feature learning. *LastN* adaptation examines the cumulative effect of levels of classifier learning. Investigating the three proposed adaptation strategies can provide novel insights into to cross-corpus generalizability.

4. Results

4.1. Experimental Settings

The experiments were conducted on two corpora recorded in very different environments: SH2-FS (Free Speech) [8], [14] and Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) [38]. The SH2-FS corpus comprises audio recordings in naturalistic environments (e.g., at home, workplace, vehicle),

along with self-reported Patient Health Questionnaire (PHQ-9) scores gathered through an interactive Android™ smartphone app. This corpus has the same training and testing partition as per [13]: 444 files (438 speakers) for training and 130 files (128 speakers) for testing. There are 74 and 23 *depressed* speakers in the training and test data partitions respectively as a result of using a PHQ-9 threshold of 10 (suggested by [39]).

The DAIC-WOZ is a laboratory-based dataset recorded during interviews with a virtual human agent via high-quality microphones with minimal background noise. Each interview produced up to 20 minutes of speech for each participant, and an accompanying clinically validate binary label indicating whether the participant was depressed or healthy. The database has 107 speakers for training and 35 speakers for testing [22]. The average speech utterance durations were 20.5 ± 10.2s for SH2-FS and 446.9 ± 227.0s for DAIC-WOZ. For both datasets, 20% of the training data were held-out as a development set for optimizing model training or adaptation.

The input to the dilated CNN, i.e. the FVTC matrix [16], consists of delayed correlations calculated from contours of short-term acoustic features for each audio recording. In this study, we employed four sets of acoustic features, namely 3 formants, 13 spectral centroid frequencies (SCF) [40], 16 MFCCs and 16 delta MFCCs (dMFCC). Unvoiced frames were dropped using voice activity detection for both corpora. All correlations were centered and scaled to unit variance, based on normalization coefficients learnt from the training set.

As for the dilated CNN structure and hyperparameters, we adopted the same architecture as per [16], namely the Adam optimizer, batch size was set to 64 for both corpora. The dilation rates n were set to 1, 3, 7, 15 in the first four parallel convolutional layers (i.e. *Conv 1* to *Conv 4*) with filter size of 15×1. *Conv-5* and *Conv-6* adopt filter size of 3×1 with a stride of 2. Batch normalization, max pooling and dropout were applied as shown in Fig. 1. The seed value was set to 0 for all experiments. Both training and adaptation was trained up to 200 epochs with early stopping based on the top average F1 score (of two classes) on the development set. Class weights were empirically set to alleviate the class imbalance issue during training. Dropout rate was fixed to 0.3 and λ for $\ell 2$ normalization was set to 0.01 unless stated otherwise.

Classification performances were evaluated using *Unweighted Averaged Recall* (UAR) ∈ [0,1] calculated for speakers, which is a standard metric to evaluate unbalanced classification problems (higher UARs are better).

4.2. Improved Results using Proposed Domain Adaptation

The first experiment evaluates the usefulness of domain adaptation by comparing three different cases, i.e., within-corpus experiments and cross-corpus experiments without and with adaptation. The learning rate was selected from {1e-3, 1e-4} for pre-training models and from {5e-3, 5e-4} for adaptation in this experiment, because the number of trainable parameters varies for adapted layers, feature types, and adaptation strategies. It was observed that, as expected, for layer-wise adaptation (i.e. less parameters), large learning rates are needed whereas for cumulative layer adaptation, smaller learning rates often gave improved performances. Models were trained and tested either on the same datasets (within-corpus) or on different datasets (cross-corpus). For the adapted cases, eight possible layer combinations (i.e. from *Conv-1* to *FC-2*) were tried, and the top results were selected.

Fig. 2 shows that adaptation consistently yielded (sometimes significant) improvements over the within-corpus and cross-corpus (without adaptation) cases. This lends support to the concept of adaptation, with the biggest gains observed for adapt/test on DAIC, although there is no clear adaptation strategy winner. Also, layer-wise adaptation performed better than or on par with *FirstN* and *LastN* on DAIC-WOZ, which is however not true for SH2-FS. This suggests that less adaptation is needed for DAIC-WOZ.

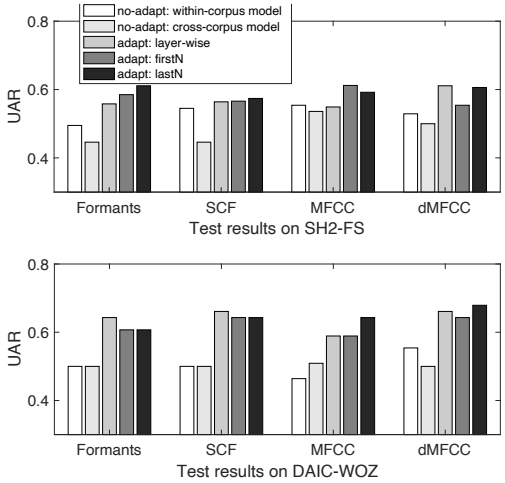

Figure 2: *Comparison of the proposed adaptation strategies against systems without domain adaptation for both corpora.*

4.3. Probing Intermediate Results of the Proposed Adaptation Strategies

This experiment evaluates the respective contributions of layer(s) during adaptation to cross-corpus generalizability (i.e. which layers need adaptation for better performance?), shown in Fig. 3. For *layer-wise*, each layer was updated separately. For *FirstN* (or *LastN*), the current layer and its preceding (or following) layers were updated.

Formants and dMFCC were chosen partly due to their strong results in Fig. 2, and partly because they represent two different cases during adaptation: dMFCC tend to be sensitive to channel variability, whereas formants are not. As shown in Fig. 3, adaptation involving *Conv-5*, *Conv-6* and *FC-1* tended to produce better performances over other layers for SH2-FS, whereas adaptation involving *Conv-1*, *Conv-3*, and *Conv-5* is beneficial for DAIC-WOZ, albeit slight variations for formant and dMFCC. This implies an interesting insight that feature

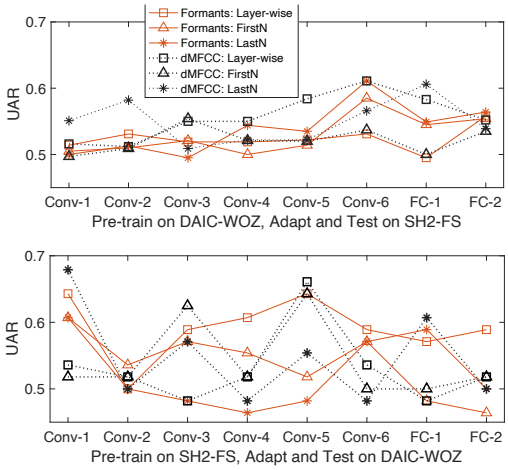

Figure 3: *UARs when adapting layer(s) using proposed adaptation strategies for both corpora.*

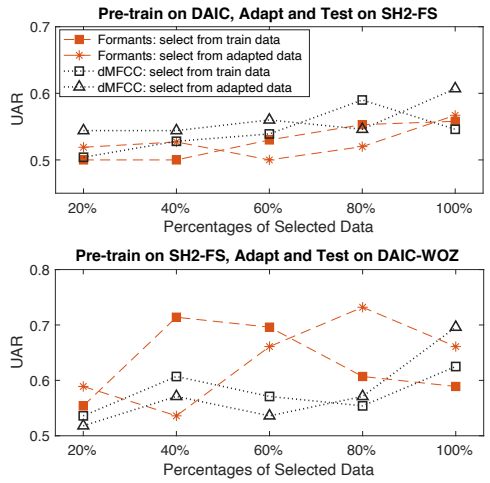

Figure 4: *Impact of data selection from 20% to 100% (i.e., using all data) of either training or adaptation partition on depression detection performances for both corpora.*

representation needs to be re-learned to accommodate the environmental changes between clean and naturalistic datasets, but the emphasis is different depending on the pre-training datasets. If pre-trained on DAIC-WOZ (clean), adaptation of *Conv-5* or *Conv-6* is sufficient for SH2-FS, whereas if pre-trained on SH2-FS (noisy), apart from *Conv-5*/*Conv-6*, detailed coordination information (i.e. from *Conv-1* to *Conv 4*) should be also adapted for DAIC-WOZ.

4.4. Impact of Amount of Data for Adaptation

An important consideration during adaptation is how much data are needed for both training and adaption. To investigate this, the amount of training or adaptation data was varied from 20% to 100% (Fig. 4). We selected the adapted system that took formants or dMFCC for pre-training and updated *Conv-5* due to their strong performances for both corpora. All the hyperparameters were identical to previous experiments.

Results in Fig. 4 show that in general, more data, either for pre-training or adaptation, tended to yield better results for both datasets, as expected. One exception was for formants, where using more than 40% of the data from SH2-FS for pre-training did not aid system performances. This is also sensible because more noisy data or poorly administered labels (from SH2-FS) are not necessarily helpful when tested on DAIC-WOZ.

4.5. Optimized Results Compared with Existing Results/Approaches

This experiment compares the optimized results with existing results on the two adopted datasets. Instead of the fixed hyperparameters used in previous experiments, for the adapted FVTC-CNN results, a grid search was performed for three hyperparameters: pre-training learning rate from {1e-3, 5e-4, 1e-4} (for better pre-trained models), adaptation learning rate from {1e-2, 5e-3, 1e-3, 5e-4, 1e-4} (which is crucial for optimized adaptation), dropout rate from {0.3, 0.4} (for regularization), alongside three adaptation strategies. Similarly, for non-adapted FVTC-CNN results (i.e. within-corpus and cross-corpus), grid search was done for three hyperparameters: batch size from {64, 128}, learning rate from {1e-2, 1e-3, 1e-4, 1e-5} (for optimization), and dropout rate from {0.2, 0.3, 0.4}.

Many interesting results can be observed from Table 1. For instance, almost all optimal results involve adapting

Table 1: *Optimized adapted results compared with non-adapted results and existing published results. F1 scores for depression (D) and healthy (H) were also presented.*

	SH2-FS		DAIC-WOZ	
	F1 *(D/H)*	UAR	F1 *(D/H)*	UAR
chance-level	0.26/0.62	0.50	0.29/0.61	0.50
eGeMAPS [13]	0.32/0.79	0.58	0.29/0.82	0.55
acoustic [8]/[22]	0.33/0.74	0.59	0.41/0.58	0.64
DepAudioNet [24]	-	-	0.52/0.70	0.77
speech landmark [13]	**0.47/0.78**	**0.73**	**0.86/0.97**	**0.91**
FVTC-CNN:				
best within-corpus	0.41/0.71	0.66$^{\{fmt\}}$	0.63/0.89	0.79$^{\{fmt\}}$
best cross-corpus	0.32/0.78	0.58$^{\{dmfcc\}}$	0.50/0.90	0.68$^{\{fmt\}}$
Adapted FVTC-CNN:				
Formants	**0.46/0.85**	**0.68$^{\{*, conv\text{-}5\}}$**	0.62/0.91	0.75$^{\{*, conv\text{-}5\}}$
SCF	0.39/0.88	0.63$^{\{\dagger, conv\text{-}4\}}$	0.57/0.89	0.73$^{\{*, conv\text{-}1\}}$
MFCC	0.37/0.76	0.62$^{\{\dagger, conv\text{-}2\}}$	**0.67/0.86**	**0.88$^{\{\dagger, fc\text{-}1\}}$**
dMFCC	0.40/0.77	0.65$^{\{\dagger, conv\text{-}2\}}$	0.67/0.91	0.80$^{\{*, conv\text{-}4\}}$

Optimized system configuration was mentioned in {}, in which '', '†', '‡' denotes layer-wise, firstN, lastN adaptation. 'fmt' means formants.*

convolutional layers, i.e. feature representation. Furthermore, similarly to Fig. 2, when adapting on DAIC-WOZ, layer-wise adaptation is preferred, whereas when adapting on SH2-FS, more layers (*firstN*) need to be adapted. This makes sense, since for the latter case, information related to the noisy conditions and handset variability needs to be learnt. An interesting insight can also be drawn from comparing formants and MFCCs in adapted FVTC-CNN: more layers need adaptation for MFCC due to its sensitivity to handset variability (i.e. †), whereas for formants, only the *Conv-5* layer needs to be adapted (i.e. *). The adapted results outperform the non-adapted cases and most existing results, confirming the effectiveness of the proposed adaptation. The adapted systems yielded relative improvements of 17.2% and 29.4% in UAR over the best non-adapted cross-corpus systems, and 3.0% and 11.4% over the best non-adapted within-corpus systems, for SH2-FS and DAIC respectively. Finally, the relatively strong results in the adapted systems for both datasets suggest that both quantity (SH2-FS) and quality (DAIC-WOZ) of data matter for pre-training models.

5. Conclusions

This study has investigated domain adaptation based on a deep learning framework to enhance cross-corpus generalizability for depression detection. Three different adaptation strategies were proposed to adapt individual or joint layers, which yielded a boost in performance over systems without adaptation. Further, contributions of intermediate CNN layers and impact of data needed for training/adaptation were studied, finding that it is important to re-learn the feature representation to accommodate environmental changes, and that adaptation benefits from more training data. Moreover, more layers need to be adapted when it comes to noisy conditions, whereas one adapted layer may be sufficient when it comes to clean conditions. As future work, this framework can be coupled with generative adversarial networks to learn environmental invariant features that are robust to noise and handset variabilities for speech-based depression detection.

6. Acknowledgements

This work was supported by Australian Research Council Linkage Project LP160101360. Julien Epps is also partly supported by Data61, CSIRO, Australia.

7. References

[1] J. Walker *et al.*, "The prevalence of depression in general hospital inpatients: a systematic review and meta-analysis of interview based studies," *Psychol. Med.*, vol. 48, no. 14, 2018.

[2] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal assessment of depression from behavioral signals," in *Handbook of Multi-Modal Multi-Sensor Interfaces*, Morgan and Claypool, 2017, pp. 113–155.

[3] T. R. Insel, "Digital phenotyping: Technology for a new science of behavior," *JAMA - J. Am. Med. Assoc.*, vol. 318, no. 13, pp. 1215–1216, 2017.

[4] 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 Commun.*, vol. 71, pp. 10–49, Jul. 2015.

[5] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, Andrew, and T. Campbell, "Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health," *Psychiatr. Rehabil. J.*, vol. 38, no. 3, pp. 218–226, 2015.

[6] K. K. Weisel, L. M. Fuhrmann, M. Berking, H. Baumeister, P. Cuijpers, and D. D. Ebert, "Standalone smartphone apps for mental health—a systematic review and meta-analysis," *npj Digit. Med.*, vol. 2, no. 1, pp. 1–10, 2019.

[7] H. Hsin *et al.*, "Transforming psychiatry into data-driven medicine with digital measurement tools," *npj Digit. Med.*, vol. 1, no. 1, pp. 1–4, 2018.

[8] Z. Huang, J. Epps, D. Joachim, and M. C. Chen, "Depression detection from short utterances via diverse smartphones in natural environmental conditions," in *INTERSPEECH*, 2018, pp. 3393–3397.

[9] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain Cogn.*, vol. 56, no. 1, pp. 30–35, 2004.

[10] J. Williamson, T. Quatieri, and B. Helfer, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on AVEC, ACM MM, Orlando, FL*, 2014.

[11] A. C. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, p. 42, 2011.

[12] S. Scherer, G. M. Lucas, J. Gratch, A. Rizzo, and L. P. Morency, "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 59–73, 2016.

[13] Z. Huang, J. Epps, and D. Joachim, "Investigation of speech landmark patterns for depression detection," *IEEE Trans. Affect. Comput.*, pp. 1–15, 2019.

[14] Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 435–448, 2020.

[15] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 4th ACM International Workshop on AVEC, ACM MM*, 2013, pp. 41–47.

[16] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated CNNs for depression detection in naturalistic environments," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6549–6553.

[17] J. Gideon, M. McInnis, and E. Mower Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2019.

[18] B. Schuller and B. Vlasenko, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, 2010.

[19] M. Shah, C. Chakrabarti, and A. Spanias, "Within and cross-corpus speech emotion recognition using latent topic model-based features," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, 2015.

[20] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, 2015, pp. 125–131.

[21] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," *Proc. 6th Int. Work. Audio/Visual Emot. Chall. - AVEC '16*, pp. 89–96, 2016.

[22] M. Valstar *et al.*, "AVEC 2016 - depression, mood, and emotion recognition workshop and challenge," pp. 3–10, 2016.

[23] F. Ringeval, M. Valstar, N. Cummins, R. Cowie, M. Schmitt, and A. Mallol-ragolta, "AVEC 2019 workshop and challenge : state-of-mind, depression with AI , and cross-cultural affect recognition," in *ACM Multimedia, AVEC '19*, 2019.

[24] 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 - AVEC '16*, 2016, pp. 35–42.

[25] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," *AVEC 2019 - Proc. 9th Int. Audio/Visual Emot. Chall. Work. co-located with MM 2019*, pp. 81–88, 2019.

[26] Z. Huang, J. Epps, and D. Joachim, "Speech landmark bigrams for depression detection from naturalistic smartphone speech," in *ICASSP*, 2019, pp. 5856–5860.

[27] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[28] D. a. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[29] M. E. Peters *et al.*, "Improving language understanding by generative pre-training," *OpenAI*, pp. 1–10, 2018.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *arXiv*, 2013, pp. 1–12.

[31] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," *Adv. Neural Inf. Process. Syst.*, no. Nips, pp. 892–900, 2016.

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

[33] M. Magill, F. Z. Qureshi, and H. W. De Haan, "Neural networks trained to solve differential equations learn general representations," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. NeurIPS, pp. 4071–4081, 2018.

[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Adv. Neural Inf. Process. Syst.*, vol. 4, no. January, pp. 3320–3328, 2014.

[35] A. Tamkin, T. Singh, D. Giovanardi, and N. Goodman, "Investigating transferability in pretrained language models," in *arXiv*, 2020.

[36] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," in *arXiv preprint arXiv:1610.01644.*, 2016.

[37] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: understanding transfer learning for medical imaging," in *arXiv:1902.07208*, 2019, no. NeurIPS.

[38] J. Gratch *et al.*, "The Distress Analysis Interview Corpus of human and computer interviews," in *LREC*, 2014, pp. 3123–3128.

[39] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001.

[40] K. Paliwal, "Spectral subband centroid features for speech recognition," in *ICASSP*, 1998, pp. 617–620.