Detecting Depression With a Temporal Context Of Speaker Embeddings

Sri Harsha Dumpala[1,2], Sebastian Rodriguez[1,2], Sheri Rempel[3], Mehri Sajjadian[4], Rudolf Uher[3,4], Sageev Oore[1,2]

[1]Faculty of Computer Science, Dalhousie University, Canada
[2]Vector Institute of AI, Canada
[3]Nova Scotia Health, Halifax, Canada
[4]Department of Psychiatry, Faculty of Medicine, Dalhousie University, Canada

Abstract

Depression detection from speech has attracted a lot of attention in recent years. However, the significance of speaker-specific information in depression detection has not yet been explored. In this work, we introduce—and analyze the significance of—speaker embeddings in a temporal context for the task of depression detection from speech. Experimental results show that the speaker embeddings provide important cues to achieve state-of-the-art performance in depression detection. We also show that combining conventional OpenSMILE and COVAREP features, which carry complementary information, with speaker embeddings further improves the depression detection performance. The significance of the temporal context in the training of deep learning models for depression detection is also analyzed in this paper.

1 Introduction

Speech is a complex signal rich in information which includes message, speaker characteristics, emotive state, etc. Speaker characteristics consist not only of identity information such as gender, age, etc., but have been shown to provide important cues about the traits of the speaker such as personality, physical state, likability and pathology (Schuller et al. 2015; Dumpala and Kopparapu 2017; Narendra and Alku 2020). Moreover, speaker-specific information was also used for emotion classification, and in detection of Alzheimer's from speech (Pappagari et al. 2020b,a). In this work, we apply—and analyze the significance of—speaker-specific information in a temporal context to the detection of depression from speech.

Major depressive disorder, also known as depression, is one of the most common mental health disorders and ranks among the health conditions responsible for most disability worldwide (Walker et al. 2018; Rehm and Shield 2019). According to World Health Organization (Organization et al. 2015), more than 300 million people (around 5% of the global population) suffer from depression, and this number is projected to further increase in the coming years. Early diagnosis of depressive symptoms is crucial in reducing the effects of this disorder.

As an attempt to aid in depression diagnosis, the problem of automatically detecting depression using speech has attracted a lot of attention (Low et al. 2010; Cummins et al. 2015; Ringeval et al. 2019; Valstar et al. 2016; Tao, Esposito, and Vinciarelli 2020; Dumpala et al. 2021b). Recently, the application of deep learning techniques have significantly boosted the performance of depression detection using speech (Tasnim and Stroulia 2019; Ma et al. 2016; Chlasta, Wołk, and Krejtz 2019; Al Hanai, Ghassemi, and Glass 2018; Huang, Epps, and Joachim 2020). Initially deep neural networks (DNNs) with fully-connected layers were trained for depression detection (Tasnim and Stroulia 2019). Later, convolutional neural networks (CNNs) and recurrent neural networks with long short-term memory (LSTM) units were shown to achieve better performance on depression detection (Chlasta, Wołk, and Krejtz 2019; Al Hanai, Ghassemi, and Glass 2018). Recently, CNN-LSTM and dilated CNNs were used for depression detection from speech to achieve state-of-the-art (SOTA) performance (Ma et al. 2016; Huang, Epps, and Joachim 2020).

Speaker-specific information, which provides important cues about speaker traits (Schuller et al. 2015), was not considered in any of the previous works on depression detection. In this work, we use speaker-specific information to train multi-kernel CNNs (CNN layer with kernels of different sizes) (Sheikh et al. 2018), and LSTM models for depression detection.

The intuition behind speaker embeddings is that similar-sounding speech segments get mapped to nearby points in the embedding space. When this is done in a text-independent manner (e.g. using contrastive objective as we describe in Section 3), then the goal is that anything said by a particular individual ought to get mapped near anything else said by that same individual. This also means that two speakers who have similar-sounding speech should get mapped near one another in the embedding space. Our hypothesis, then, is that if the speech of people suffering from depression tends to have certain consistent (similar) characteristics, then their speech will, also, get mapped to nearby positions in the embedding space. If this is true, then this would allow us—as we will demonstrate in this paper—to first learn speaker embeddings from huge and unlabelled (but accessible) datasets, and then leverage this information to detect depression by training on small (and hard to obtain) clinical datasets that have been collected and labelled specifically for the context of depression analysis. We hypothesize and

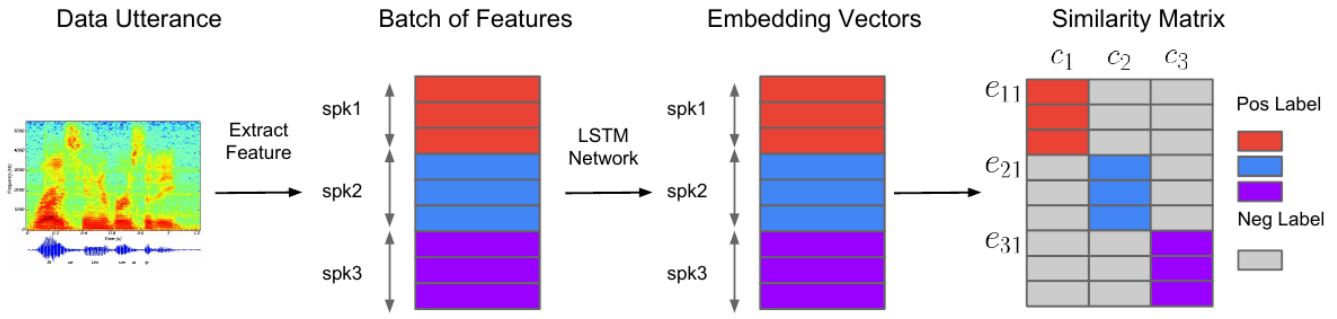

Figure 1: Overview of the generalized end-to-end loss speaker verification system (Figure taken from (Wan et al. 2018)). Different colors indicate utterances/embeddings from different speakers.

show that this transfer can work effectively even though the two sources of data (i.e., huge unlabelled dataset, and the small clinical dataset) do not have any speakers in common.

The main contributions of this work are as follows:

- Introduce the application of generalized speaker embeddings for depression detection using speech and analyze their significance for this task.
- Demonstrate that we can leverage a large unlabelled speech dataset to learn embeddings which can be used to obtain state-of-the-art performance on depression detection using much smaller medically-labelled datasets.
- Analyze the significance of temporal context, i.e., number of contiguous segments to be considered, in training deep learning models for depression detection.

The rest of the paper is organized as follows. Section 2 provides a brief review of the related work. In Section 3, we describe our proposed approach for depression detection, and provide background on speaker embedding extraction. We provide details of the datasets used for our analysis in Section 4. Experimental setup and results are discussed in Section 5. Summary of this work is given in Section 6.

2 Related Work

Acoustic Representations for Depression Detection: Depression is shown to degrade cognitive planning and psychomotor functioning, thus affecting the human speech production mechanism (Cummins et al. 2015). These effects manifest as variations in the speech voice quality (Williamson et al. 2014) and several features have been proposed to capture these variations in speech for depression detection. Spectral features such as formants and mel-frequency cepstral coefficients (MFCCs), prosodic features such as F_0, jitter, shimmer and glottal features were initially used for depression detection (Low et al. 2010; Cummins et al. 2011; Simantiraki et al. 2017). Spectral, prosodic and other voice quality related features extracted using OpenSMILE (Eyben, Wöllmer, and Schuller 2010) and COVAREP (Degottex et al. 2014) toolkits were also used for depression analysis (Valstar et al. 2016; Al Hanai, Ghassemi, and Glass 2018). Further, features developed based on speech articulation such as vocal tract coordination features were analyzed for depression detection (Williamson et al. 2014; Huang, Epps, and Joachim 2020; Seneviratne et al. 2020). Recently, sentiment and emotion embeddings, representing non-verbal characteristics of speech, were used for depression severity estimation (Dumpala et al. 2021a).

To the best of our knowledge, no other studies that we know of have explored the use of speaker-specific information for depression detection. In this work, we consider using speaker embeddings, a representation of speaker-specific information, for depression detection.

Speaker Embeddings: Speaker embeddings refer to a low-dimensional representation of the speaker-specific characteristics that exist in the speech signal (Snyder et al. 2016, 2018) and can be designed to be relatively *independent* of *what* the speaker is saying. Speaker representations were initially based on i-vectors, with a probabilistic linear discriminant analysis (PLDA) back-end (Dehak et al. 2010). More recently, two distinct end-to-end deep neural network based approaches were used for speaker verification, and both approaches obtained (comparable) state-of-the-art performance (Snyder et al. 2018; Wan et al. 2018). In Snyder et al. (2018), speaker embeddings, also referred to as x-vectors, were extracted from a time-delay deep neural network trained for the task of speaker verification. In contrast, Wan et al. (2018) extracted speaker embeddings, also referred to as d-vectors, from an end-to-end LSTM network trained for speaker verification. In this paper, we use the generalized end-to-end text-independent speaker verification system (shown in Figure 1) proposed in Wan et al. (2018).

Temporal Context in Depression Detection: A few studies have analyzed the effect of the total duration of the audio recording on the depression detection performance (Yang et al. 2016; Pampouchidou et al. 2016; Rutowski et al. 2019). These works have shown that longer the duration, better the performance. In Yang et al. (2016); Pampouchidou et al. (2016), the analysis was performed by considering multiple modalities i.e, audio, visual and text, whereas in Rutowski et al. (2019), automatic speech-to-text transcriptions were used to analyze the effect of duration on depression detection performance. In this work, we use the acoustic features extracted from speech to analyze the effect of varying the number of contiguous speech segments on the per-

formance of LSTM and CNN models trained for depression detection. In particular, after establishing the significance of speaker embeddings for depression detection, we analyze how the temporal variation in speaker embeddings across different context lengths will affect the depression detection performance.

3 Proposed Approach

Extraction of Speaker Embeddings

In this work, we obtain speaker embeddings from speech using the generalized end-to-end (GE2E) speaker verification model proposed in Wan et al. (2018). We provide below a brief review of the GE2E approach (see Figure 1). GE2E training is based on processing a large number of utterances at once, in the form of a batch that contains N speakers, and M utterances from each speaker. Each feature vector sequence x_{ji} ($1 \leq j \leq N$ and $1 \leq i \leq M$) represents the feature sequence (frame-level features) extracted from the i^{th} utterance of j^{th} speaker. The length of the feature sequence x_{ji} of the audio samples varied from 140 to 180 frames. These frame-level features extracted from each utterance x_{ji} were fed into a deep LSTM network with 3 LSTM layers (with 256 units each) followed by a linear layer (with 256 units). The linear layer performs an affine transformation on the last frame response of the LSTM layers. The output of the linear layer of the network is denoted as $f(x_{ji};\theta)$ where θ represents parameters of the entire neural network. The embedding vector (also known as d-vector) is defined as the $L2$ normalization of the final layer output:

$$e_{ji} = \frac{f(x_{ji},\theta)}{\|f(x_{ji},\theta)\|},$$

where e_{ji} represents the embedding vector obtained for the i^{th} utterance of j^{th} speaker. In Figure 1, c_j refers to the centroid (represents the voice print) of the j^{th} speaker obtained by computing the mean of all the embedding vectors $[e_{j1}, e_{j2}, \cdots e_{jM}]$ corresponding to j^{th} speaker. A similarity matrix S is computed for each batch, with $N \times M$ rows and N columns. An element $S_{ji,k}$ in the similarity matrix is defined as the scaled cosine similarity between each embeddings vector e_{ji} and all the centroids c_k ($1 \leq j,k \leq N$ and $1 \leq i \leq M$):

$$S_{ji,k} = w \cdot cos(e_{ji}, c_k) + b,$$

where w and b are learnable parameters. For the similarity matrix shown in Figure 1, we want the similarity values of colored areas (Pos Label) to be large, and the values of gray areas (Neg Label) to be small. To achieve this objective, the model was trained using softmax on $S_{ji,k}$, which outputs 1 if $j = k$, otherwise outputs 0. The softmax loss on embeddings vector e_{ji} can be defined as:

$$L(e_{ji}) = S_{ji,j} - log \sum_{k=1}^{N} exp(S_{ji,k}).$$

This loss function enables the network to learn parameters such that embeddings vectors corresponding to j_{th} speaker are pulled close to the centroid c_j and at the same time pushed away from other centroids corresponding to other speakers.

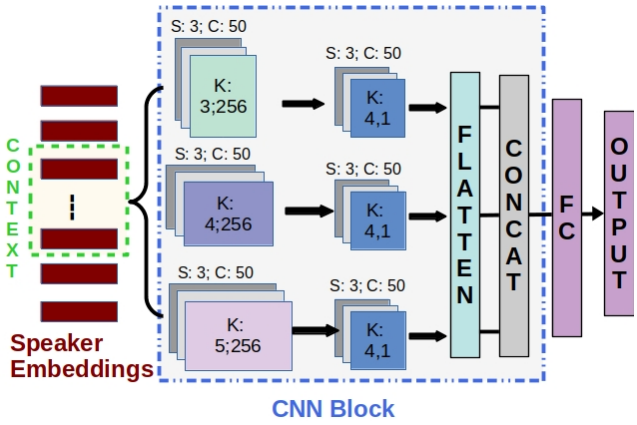

Figure 2: Network for depression detection using speaker embeddings as input. The same network is used for OpenSMILE and COVAREP features. FC refers to a fully-connected layer.

In this work, we pre-trained the GE2E network on the task of speaker verification by consolidating 3 different datasets i.e., LibriSpeech (Panayotov et al. 2015), VoxCeleb1 and VoxCeleb2 (Nagrani, Chung, and Zisserman 2017) with 1166 speakers, 1211 speakers and 5994 speakers, respectively. The three pre-training datasets consolidated together consist of a total of around 1.4 million utterances (each utterance is of 5 to 10 seconds in duration), totalling about 3300 hours of audio, collected from 8371 different speakers. Each batch consisted of N = 64 speakers and M = 10 utterances per speaker. 40-dimensional MFCCs extracted using a window of size 30 msec and a step size of 10 msec are used as the input features. In training, we randomly selected 160 contiguous frames for each sample. Note that *we did not train the GE2E network on the depression datasets (i.e., DAIC-WoZ and FORBOW datasets).*

We then used this GE2E model to extract speaker embeddings at segment-level for each of the DAIC-WoZ and FORBOW datasets. Each segment is represented using a speaker embedding of dimension 256. Finally, we use these speaker embeddings to train and test the LSTM and CNN based models (explained in Section 3) for depression detection.

Speaker Embeddings for Depression Detection

We explore the use of both CNN (shown in Figure 2) and LSTM networks for depression detection when the speaker embeddings are provided as input.

CNN for Depression Detection (CNN$_D$): A CNN with multiple kernels (Sheikh et al. 2018), as shown in Figure 2, is used for depression detection from the extracted speaker embeddings. The first convolutional layer consists of 3 different kernels with sizes $(3, D)$, $(4, D)$ and $(5, D)$, respectively. Here, D refers to the length of the input feature vector. For instance, in Figure 2, D is 256 for CNN$_D$ network trained with speaker embedding (length of

speaker embedding is 256). Similarly, for CNN_D trained with OpenSMILE features, D is 384, and for CNN_D trained with COVAREP features, D is 444. In other words, the kernel dimensions for the first layer of the CNN_D network trained with OpenSMILE features are $(3, 384)$, $(4, 384)$ and $(5, 384)$. Similarly, for CNN_D trained with COVAREP features, the kernel dimensions are $(3, 444)$, $(4, 444)$ and $(5, 444)$. Each kernel consists of 50 channels. In the second convolutional layer, all kernels are of size 4 with 50 channels in each kernel. Outputs from each kernel of the second convolutional layer are flattened and then concatenated before passing through a fully-connected (FC) layer with 100 units, and then through an output softmax layer with 2 units.

LSTM for Depression Detection (LSTM_D): Depression detection from the extracted speaker embeddings is also performed by considering an LSTM network. The LSTM network is same as the CNN_D network shown in Figure 2, but the CNN block is replaced by an LSTM block, consisting of 2 LSTM layers with 128 units each. The output of the LSTM block for the last timestep is passed through the FC layer with 100 units, and then through an output softmax layer with 2 units for obtaining the final decision.

Baseline FC Network for Depression Detection (DNN_D): We considered a fully-connected deep neural network (DNN) for comparison. This DNN has 3 hidden layers with 128, 64 and 128 ReLU units, respectively, and a softmax output layer with 2 units for obtaining the final decision.

Further, we extracted COVAREP (Degottex et al. 2014; Al Hanai, Ghassemi, and Glass 2018) and OpenSMILE (Eyben, Wöllmer, and Schuller 2010) features for performance comparison with speaker embeddings. We extracted COVAREP and OpenSMILE features at segment-level to train and test the CNN_D, LSTM_D and DNN_D networks. 384-dimensional OpenSMILE features representing each segment were obtained by using the $IS09$ configuration from OpenSMILE toolkit. We obtained segment-level COVAREP features (444-dimensional) by computing the higher-order statistics (mean, maximum, minimum, standard deviation, skew, and kurtosis) of the 74-dimensional frame-level features (frame-size of 20 msec and frame-shift of 10 msec).

Combined Embeddings for Depression (CE_D): We also try combining speaker embeddings with each of OpenSMILE or COVAREP features, respectively (Figure 3), for depression detection. As shown in Figure 3, the proposed network consists of two branches, one for speaker embeddings and the other for OpenSMILE or COVAREP features. In each branch, the input features are passed through an LSTM (CE_{DL}) or CNN (CE_{DC}) block and then through a fully-connected (FC) layer (100 units). The outputs of the FC layer of each branch are combined, and then passed through an output layer to get the final decision. Various techniques (summation, dot product, concatenation and average) were considered for combining the outputs of the two

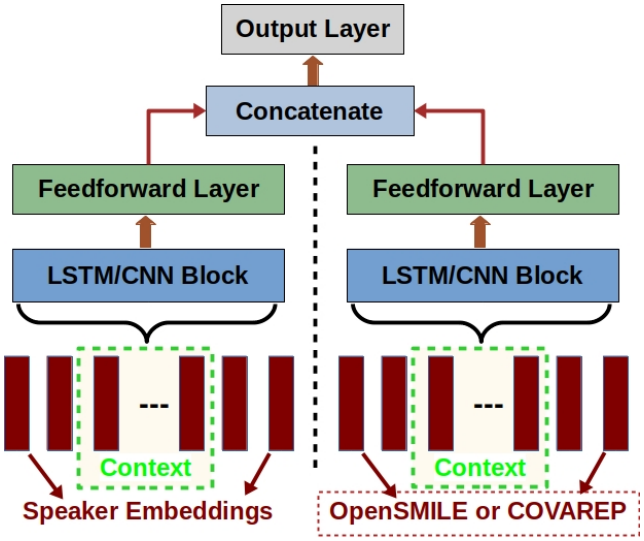

Figure 3: Network combining speaker embeddings, and OpenSMILE or COVAREP features for depression detection.

branches. We found that combining the outputs of the two branches (using any of the combining techniques) resulted in better/similar performance when compared to the case of using only speaker embeddings. Out of all combining techniques, dot product gave the best results. All the results reported in this paper are based on dot product.

The context in Figures 2 and 3 refers to the number of contiguous segments in an audio recording considered to train and test the models. we experiment with temporal contexts of different length to analyze the optimal number of contiguous segments required to train the CNN_D and LSTM_D models for better performance (see Section 5). Note that even though the networks are trained and tested at segment-level with different contexts, the final accuracy is based on the prediction for the entire audio file. Majority voting is performed on the segment-level decisions to obtain the final decision i.e., depressed or not-depressed.

Training Details: All the networks were trained using Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$) with an initial learning rate of 0.0005. Dropout rate of 0.3, 0.4 and 0.3 was used for the CNN block, LSTM block and FC layers, respectively to avoid model over-fitting. ReLU activation was used for all the CNN, LSTM and FC layers. Softmax activation for the output layer. All networks were trained for 50 epochs with a batch size of 128. Negative log-likelihood (NLL) loss functions was used to train models. Class weights were set based on the distribution of samples in the train set to alleviate the class imbalance issue during training. The same temporal context (number of contiguous segments in a sample) was maintained in the train, validation and testing phase.

4 Clinical Data

In this work, we considered two different depression datasets i.e., DAIC-Woz (corpus of clinical interviews) and

Table 1: Depression detection performance in terms of F_1 and Accuracy (Acc.), when speaker embeddings are considered.

	Model	Context	F_{1D}	F_{1H}	Acc.
DAIC	DNN_D	1	0.32	0.74	0.63
	CNN_D	20	0.42	0.77	0.68
	$LSTM_D$	20	**0.44**	**0.78**	**0.69**
FORB.	DNN_D	1	0.28	0.74	0.65
	CNN_D	16	0.31	0.79	0.70
	$LSTM_D$	16	**0.34**	**0.79**	**0.71**

FORBOW (spontaneous speech corpus obtained in a clinical setting) for analysis.

DAIC-WoZ: The DAIC-WoZ dataset contains a set of 189 clinical interviews collected from 189 individuals. Each interview was conducted between a client and a virtual agent controlled by a human interviewer placed in another location (Gratch et al. 2014). Each audio file was labeled with a PHQ-8 (Patient Health Questionnaire) score which is in the range of $0 - 24$ to denote the severity of depression. Audio files with depression score (PHQ-8) 10 or above are considered to be depressed, and those audio files with depression scores below 10 are considered as healthy/non-depressed. Out of the 189 recordings, 132 are collected from healthy subjects and remaining 57 are collected from depressed subjects. The audio recordings of the dataset are divided into train, validation and test sets, consisting of 107 (76 healthy and 31 depressed), 35 (23 healthy and 12 depressed), 47 (33 healthy and 14 depressed) audio samples, respectively (adopted same partitions as in (Valstar et al. 2016)). The train, validation and test splits do not have any overlapped speakers. Timestamps were provided to each response of the client to the interviewer questions. In this work, each recording is divided into non-overlapping segments of at least 5 seconds duration. Multiple contiguous responses are combined to form a single segment, if the duration of a response is less than 5 seconds leading to a total of 13386 segments (train set: 7255, validation set: 2548 and test set: 3583 segments).

FORBOW Dataset: We also analyze speech data collected as part of the FORBOW[1] medical research project (Uher et al. 2014). Speech samples were collected from 514 subjects (390 mothers and 124 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 (Montgomery and Åsberg 1979)[2]. The range of MADRS scores in this database is $0 - 21$. Audio files with depression score (MADRS) 10 or above are considered to be depressed, and those audio files with depressive scores below 10 are considered as non-depressed. Out of the 514 recordings, 403 are collected from healthy subjects and the remaining 111 are collected from depressed subjects. The dataset is divided into train, validation and test sets, consisting of 353 (279 healthy and 74 depressed), 61 (47 healthy and 14 depressed), 100 (77 healthy and 33 depressed) samples, respectively. There is no overlap in speakers between the train, validation and test splits. Each audio recording is divided into non-overlapping segments of 5 second duration for training and testing the machine learning models. A total of 25772 segments (train set: 17524, validation set: 3189 and test set: 5059 segments) are obtained.

For both datasets, the depression label (i.e., depressed or healthy) assigned to a segment obtained from an audio recording is the same as the depression label of the entire audio recording.

5 Experimental Results

Depression detection performance scores when speaker embeddings are used to train DNN_D, CNN_D and $LSTM_D$ models are given in Table 1. It can be observed from Table 1 that the LSTM and CNN models achieve better performance when compared to DNN on both DAIC-WoZ (DAIC) and FORBOW (FORB.) datasets. F_{1D} and F_{1H} are F_1 scores of depressed and healthy classes, respectively, and *Acc.* refers to the weighted accuracy of the two classes i.e., depressed and non-depressed (healthy).

Table 2 shows the depression detection performance when speaker embeddings are combined with each of OpenSMILE (Spk-Emb, OS) or COVAREP (Spk-Emb, COV) features, respectively. It can be observed from Tables 1 and 2 that the models trained on speaker embeddings outperform the models trained on COVAREP or OpenSMILE features for both DAIC-WoZ and FORBOW datasets. It can also be observed that combining speaker embeddings with OpenSMILE or COVAREP features further improves the depression detection performance. This shows that the speaker embeddings carry complementary information when compared to OpenSMILE or COVAREP features. Moreover, the $LSTM_D$ and CNN_D outperformed the DNN_D in all conditions, with the $LSTM_D$ performing better or similar to the CNN_D models. In Table 1 and Table 2, context refers to the number of contiguous segments for which we obtained the best performance for a model trained on the corresponding feature.

We compare the performance of our proposed approach with previous state-of-the-art (SOTA) approaches for depression detection (see Table 3). In sequence (Al Hanai, Ghassemi, and Glass 2018), LSTM networks were trained with Covarep features for depression detection. In eGeMAPS (Huang, Epps, and Joachim 2019), CNN models were trained using OpenSMILE features. In FVTC-MFCC (Huang, Epps, and Joachim 2020), channel-delayed correlations of MFCCs were used to train dilated CNN models. In FVTC-FMT (Huang, Epps, and Joachim 2020), channel-delayed correlations of formant frequencies were used to train dilated CNN models. None of these approaches

[1]Families Overcoming Risks and Building Opportunities for Well Being

[2]Both MADRS and PHQ-8 are used in clinical practise; MADRS is considered more reliable but also more time-consuming as it is measured by a clinician whereas PHQ-8 is self-reported by the patient.

Table 2: Depression detection performance when speaker embeddings (Spk-Emb) are combined with COVAREP (COV) and OpenSMILE (OS) features. CE_{DD}, CE_{DC} and CE_{DL} refer to CE_D with DNN, CNN and LSTM blocks, respectively

	COVAREP					(Spk-Emb, COV)				
DAIC-WoZ		context	F_{1D}	F_{1H}	Acc.		context	F_{1D}	F_{1H}	Acc.
	DNN_D	1	0.31	0.64	0.56	CE_{DD}	1	0.32	0.74	0.63
	CNN_D	18	0.36	**0.71**	**0.61**	CE_{DC}	20	0.43	0.78	0.69
	$LSTM_D$	18	**0.37**	0.69	0.60	CE_{DL}	20	**0.46**	**0.78**	**0.70**
	OpenSMILE					(Spk-Emb, OS)				
		context	F_{1D}	F_{1H}	Acc.		context	F_{1D}	F_{1H}	Acc.
	DNN_D	1	0.31	0.70	0.59	CE_{DD}	1	0.34	0.76	0.65
	CNN_D	20	0.35	0.73	0.63	CE_{DC}	20	0.48	0.80	0.72
	$LSTM_D$	20	**0.36**	**0.74**	**0.64**	CE_{DL}	20	**0.50**	**0.82**	**0.74**
FORBOW	COVAREP					(Spk-Emb, COV)				
		context	F_{1D}	F_{1H}	Acc.		context	F_{1D}	F_{1H}	Acc.
	DNN_D	1	0.29	0.67	0.59	CE_{DD}	1	0.30	0.75	0.67
	CNN_D	16	0.31	**0.69**	**0.62**	CE_{DC}	16	0.34	**0.80**	0.72
	$LSTM_D$	16	**0.34**	0.68	**0.62**	CE_{DL}	16	**0.35**	**0.80**	**0.72**
	OpenSMILE					(Spk-Emb, OS)				
		context	F_{1D}	F_{1H}	Acc.		context	F_{1D}	F_{1H}	Acc.
	DNN_D	1	0.24	0.72	0.62	CE_{DD}	1	0.35	0.77	0.69
	CNN_D	16	**0.28**	0.75	**0.66**	CE_{DC}	16	0.39	0.82	0.74
	$LSTM_D$	16	0.26	**0.76**	**0.66**	CE_{DL}	16	**0.42**	**0.83**	**0.76**

considered speaker-specific features for depression detection. Table 3 shows that the models trained on speaker embeddings perform better than (or at least comparable to) the SOTA approaches for depression detection using speech on DAIC-WOZ and FORBOW datasets. It can also be observed that the depression detection performances obtained by combining speaker embeddings with the OpenSMILE features (Spk-Emb, OS) outperform the SOTA approaches.

In Table 3, for DAIC-WOZ, we also provide results on the validation set simply for comparison purposes: as this data was originally released as part of a competition, and Al Hanai, Ghassemi, and Glass (2018) and Huang, Epps, and Joachim (2019) were only able to provide results on the validation set in their original paper. Since then, the test set has been made available as well, so we also provide results on the test set. For FORBOW, as there are no other papers that use this dataset, we provide results only on the test set.

Comparison with Other Pre-trained Embeddings

We compare the performance of the proposed speaker embeddings with embeddings extracted using other pre-training techniques i.e., vq- wav2vec, wav2vec 2.0, TRILL and Mockingjay (Baevski, Schneider, and Auli 2019; Baevski et al. 2020; Shor et al. 2020; Liu et al. 2020). In these experiments, we trained the CNN_D and $LSTM_D$ networks with the speech-based embeddings extracted from the different pre-trained models. In Table 4, we report results using the $LSTM_D$ networks which performed better than the CNN_D networks. Table 4 shows that the speaker-specific embeddings perform better than the speech-based embeddings extracted using other pre-trained models.

We also analyzed the effectiveness of the extracted speaker embeddings for the task of speaker classification. DAIC-WOZ and FORBOW datasets consist of audio recordings corresponding to 189 and 517 speakers, respectively. For each speaker, 25 and 15 non-overlapping segments were randomly selected to form the train and test sets for that speaker, respectively. A logistic regression classifier (with no hidden layers) was trained for the task of speaker classification (189 and 517 class classification for DAIC-WoZ and FORBOW datasets, respectively). On the test sets, equal error rates (EER) of 1.29 and 1.69 are obtained for DAIC-WoZ and FORBOW datasets, respectively. These low EER values show that the extracted speaker embeddings, used in this work, carry speaker-specific information.

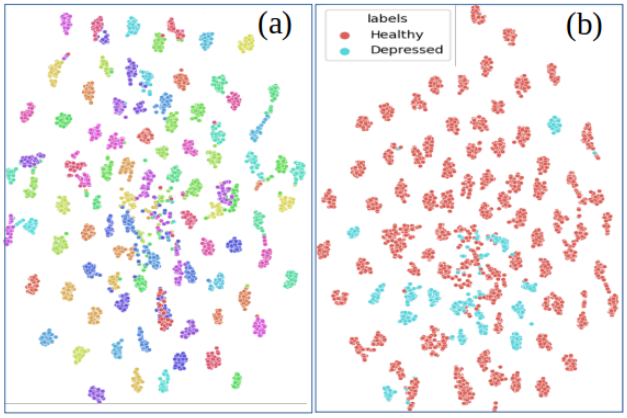

Figure 4: t-SNE plots to visualize (a) the generalized speaker embeddings in terms of speaker variations, (b) the generalized speaker embeddings based on the depression labels.

To visualize the variation in speaker embeddings, we generated the t-SNE plots (Van der Maaten and Hinton 2008)

Table 3: Depression detection performances comparing proposed approaches with state-of-the-art approaches. CE_{DC} and CE_{DL} refer to combined embeddings with CNN and LSTM blocks, respectively. Spk-Emb, COV and OS refer to speaker embeddings, COVAREP and OpenSMILE features, respectively.

	Approach	Validation set			Test set		
		F_{1D}	F_{1H}	Acc.	F_{1D}	F_{1H}	Acc.
DAIC-WoZ	Sequence (Al Hanai, Ghassemi, and Glass 2018)	0.34	0.76	0.65	0.37	0.69	0.60
	eGeMAPS (Huang, Epps, and Joachim 2019)	0.32	0.79	0.67	0.31	0.70	0.59
	FVTC-MFCC (Huang, Epps, and Joachim 2020)	0.42	0.82	0.73	0.38	0.78	0.67
	FVTC-FMT (Huang, Epps, and Joachim 2020)	0.44	0.81	0.73	0.41	0.78	0.68
	CNN_D(Spk-Emb) –from Table 1	0.41	0.82	0.73	0.42	0.77	0.68
	LSTM_D(Spk-Emb) –from Table 1	0.43	0.82	0.73	0.44	0.78	0.69
	CE_{DC}(Spk-Emb, COV) –from Table 2	0.44	0.83	0.74	0.43	0.78	0.69
	CE_{DL}(Spk-Emb, COV) –from Table 2	0.46	0.84	0.75	0.46	0.78	0.70
	CE_{DC}(Spk-Emb, OS) –from Table 2	0.50	0.85	0.77	0.48	0.80	0.72
	CE_{DL}(Spk-Emb, OS) –from Table 2	**0.51**	**0.86**	**0.78**	**0.50**	**0.82**	**0.74**
FORBOW	Sequence (Al Hanai, Ghassemi, and Glass 2018)	–	–	–	0.34	0.68	0.62
	eGeMAPS (Huang, Epps, and Joachim 2019)	–	–	–	0.25	0.73	0.63
	FVTC-MFCC (Huang, Epps, and Joachim 2020)	–	–	–	0.28	0.76	0.67
	FVTC-FMT (Huang, Epps, and Joachim 2020)	–	–	–	0.33	0.77	0.69
	CNN_D(Spk-Emb) –from Table 1	–	–	–	0.31	0.79	0.70
	LSTM_D(Spk-Emb) –from Table 1	–	–	–	0.34	0.79	0.71
	CE_{DC}(Spk-Emb, COV) –from Table 2	–	–	–	0.34	0.80	0.72
	CE_{DL}(Spk-Emb, COV) –from Table 2	–	–	–	0.35	0.80	0.72
	CE_{DC}(Spk-Emb, OS) –from Table 2	–	–	–	0.39	0.82	0.74
	CE_{DL}(Spk-Emb, OS) –from Table 2	–	–	–	**0.42**	**0.83**	**0.76**

Table 4: Comparing the performance of speaker embeddings with other pre-trained embeddings for depression detection.

	Model	F_{1D}	F_{1H}	Acc.
DAIC	TRILL	0.34	0.76	0.66
	vq-wav2vec	0.31	0.70	0.61
	wav2vec-2.0	0.36	0.73	0.63
	Mockingjay	0.29	0.68	0.58
	Spk-Emb	0.44	0.78	0.69
FORBOW	TRILL	0.30	0.75	0.67
	vq-wav2vec	0.27	0.72	0.63
	wav2vec-2.0	0.28	0.73	0.64
	Mockingjay	0.25	0.69	0.60
	Spk-Emb	0.34	0.79	0.71

for the speaker embeddings extracted from the test set (100 recordings) of the FORBOW dataset. Figure 4 shows the t-SNE plots obtained from the generalized speaker embeddings in terms of (a) the speaker label and (b) the depression label. It can be visualized from the t-SNE plots (Figure 4 (a)) that the speaker embeddings extracted from speech collected from different speakers form well separated clusters by maintaining the intra-class and inter-class variations i.e., speaker embeddings of samples corresponding to the same speaker are very close and mostly farther apart from the speaker embeddings obtained from other speakers. Figure 4(b) shows the t-SNE plots of the speaker embeddings when labeled in terms of depressed and healthy speakers. It can be observed from Figure 4(b) that the generalized speaker embeddings obtained from most of the depressed speakers are nearby which imply that the speakers with depression might share some common speaker characteristics.

Temporal Context in Depression Detection

Figure 5 shows the depression detection performance on the DAIC-Woz (DAIC) and FORBOW (FORB) datasets when different temporal contexts are considered. Two different input configurations are used for each data set: one uses only Speaker Embeddings, and the other combines Speaker Embeddings and OpenSMILE (Spk-Emb, OS) features. Note that it does not make sense to use a context longer than the shortest speech samples in the test set, which are 16 and 20 segments in FORBOW and DAIC respectively. Thus our temporal contexts range from roughly 20 seconds (4 contiguous segments) to 80 or 100 seconds (16 or 20 contiguous segments).

For all cases, as we increase the temporal context, the depression detection performance tends to improve until saturation. For example, Figure 5 shows that for the CE_{DL} model trained using combined speaker embeddings and OpenSMILE on FORBOW dataset (i.e., FORB:Spk-Emb, OS), as we increase the temporal context up to 16 segments, the performance of the CE_{DL} improves to an accuracy of 0.76. This indicates that the temporal relationship existing in the features provide important cues for depression detection.

An important area of future work, as larger data sets become available, will be to investigate the relationship of overall duration of the audio samples and the context size at which performance saturates.

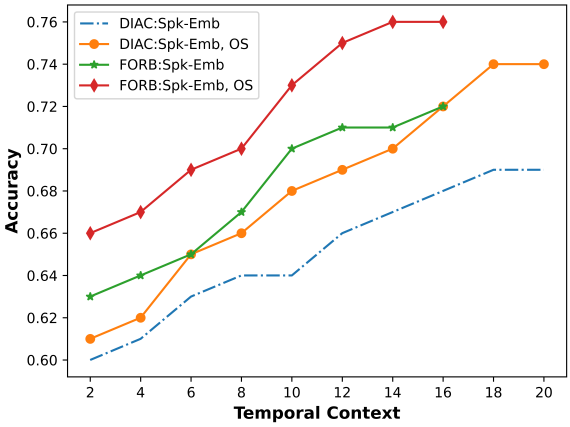

Figure 5: Performance (Accuracy) of the LSTM_D (Spk-Emb) and CE_{DL} (Spk-Emb, OS) for depression detection when the length of the context is varied from 4 up to the entire length of the shortest example in the test set, i.e. 16 and 20 respectively.

6 Summary

In this work we train a speaker embedding network on standard large datasets and then use two small clinical datasets to show that the resulting embeddings can then be used to detect depression from speech. In particular, when we combine these embeddings with OpenSMILE speech features, we achieve state-of-the-art performance on the depression detection task. Finally, our experimental results show that as we increase the temporal context (i.e., number of contiguous segments considered to train and test deep learning models), the depression detection performance improves.

References

Al Hanai, T.; Ghassemi, M. M.; and Glass, J. R. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews. In *Interspeech*, 1716–1720.

Baevski, A.; Schneider, S.; and Auli, M. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Chlasta, K.; Wołk, K.; and Krejtz, I. 2019. Automated speech-based screening of depression using deep convolutional neural networks. *Procedia Computer Science*, 164: 618–628.

Cummins, N.; Epps, J.; Breakspear, M.; and Goecke, R. 2011. An investigation of depressed speech detection: Features and normalization. In *Interspeech*.

Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; and Quatieri, T. F. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *ICASSP*, 960–964. IEEE.

Dehak, N.; Kenny, P. J.; Dehak, R.; Dumouchel, P.; and Ouellet, P. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4): 788–798.

Dumpala, S. H.; and Kopparapu, S. K. 2017. Improved speaker recognition system for stressed speech using deep neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 1257–1264. IEEE.

Dumpala, S. H.; Rempel, S.; Dikaios, K.; Sajjadian, M.; Uher, R.; and Oore, S. 2021a. Estimating Severity of Depression From Acoustic Features and Embeddings of Natural Speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7278–7282. IEEE.

Dumpala, S. H.; Uher, R.; Matwin, S.; Kiefte, M.; and Oore, S. 2021b. Sine-Wave Speech and Privacy-Preserving Depression Detection. In *Proc. SMM21, Workshop on Speech, Music and Mind 2021*, 11–15.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. ACM conference on Multimedia*, 1459–1462.

Gratch, J.; Artstein, R.; Lucas, G. M.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, 3123–3128.

Huang, Z.; Epps, J.; and Joachim, D. 2019. Investigation of speech landmark patterns for depression detection. *IEEE Transactions on Affective Computing*.

Huang, Z.; Epps, J.; and Joachim, D. 2020. Exploiting vocal tract coordination using dilated CNNs for depression detection in naturalistic environments. In *ICASSP*, 6549–6553. IEEE.

Liu, A. T.; Yang, S.-w.; Chi, P.-H.; Hsu, P.-c.; and Lee, H.-y. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6419–6423. IEEE.

Low, L. A.; Maddage, N. C.; Lech, M.; Sheeber, L.; and Allen, N. 2010. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In *ICASSP*. IEEE.

Ma, X.; Yang, H.; Chen, Q.; Huang, D.; and Wang, Y. 2016. Depaudionet: An efficient deep model for audio based depression classification. In *workshop on Audio/visual emotion challenge*.

Montgomery, S. A.; and Åsberg, M. 1979. A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Narendra, N.; and Alku, P. 2020. Glottal source information for pathological voice detection. *IEEE Access*, 8.

Organization, W. H.; et al. 2015. The European mental health action plan 2013–2020. *Copenhagen: World Health Organization*, 17.

Pampouchidou, A.; Simantiraki, O.; Fazlollahi, A.; Pediaditis, M.; et al. 2016. Depression assessment by fusing high and low level features from audio, video, and text. In *Proc. ACM workshop on Audio/visual emotion challenge*, 27–34.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 5206–5210. IEEE.

Pappagari, R.; Cho, J.; Moro-Velazquez, L.; and Dehak, N. 2020a. Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity. *Proc. Interspeech 2020*, 2177–2181.

Pappagari, R.; Wang, T.; Villalba, J.; Chen, N.; and Dehak, N. 2020b. x-vectors meet emotions: A study on dependencies between emotion and speaker recognition. In *ICASSP*. IEEE.

Rehm, J.; and Shield, K. D. 2019. Global burden of disease and the impact of mental and addictive disorders. *Current psychiatry reports*, 21(2): 10.

Ringeval, F.; Schuller, B.; Valstar, M.; Cummins, N.; Cowie, R.; Tavabi, L.; Schmitt, M.; Alisamir, S.; et al. 2019. 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*, 3–12.

Rutowski, T.; Harati, A.; Lu, Y.; and Shriberg, E. 2019. Optimizing Speech-Input Length for Speaker-Independent Depression Classification. In *INTERSPEECH*, 3023–3027.

Schuller, B.; Steidl, S.; Batliner, A.; Nöth, E.; Vinciarelli, A.; et al. 2015. A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. *Computer speech & language*, 29(1): 100–131.

Seneviratne, N.; Williamson, J. R.; Lammert, A. C.; Quatieri, T. F.; and Espy-Wilson, C. 2020. Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression. In *Proc. Interspeech*, volume 2020.

Sheikh, I.; Dumpala, S. H.; Chakraborty, R.; and Kopparapu, S. K. 2018. Sentiment analysis using imperfect views from spoken language and acoustic modalities. In *Proc. Grand Challenge and Workshop on Human Multimodal Language*, 35–39.

Shor, J.; Jansen, A.; Maor, R.; Lang, O.; Tuval, O.; Quitry, F. d. C.; Tagliasacchi, M.; Shavitt, I.; Emanuel, D.; and Haviv, Y. 2020. Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*.

Simantiraki, O.; Charonyktakis, P.; Pampouchidou, A.; Tsiknakis, M.; and Cooke, M. 2017. Glottal Source Features for Automatic Speech-Based Depression Assessment. In *INTERSPEECH*, 2700–2704.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, 5329–5333. IEEE.

Snyder, D.; Ghahremani, P.; Povey, D.; Garcia-Romero, D.; Carmiel, Y.; and Khudanpur, S. 2016. Deep neural network-based speaker embeddings for end-to-end speaker verification. In *SLT Workshop*, 165–170. IEEE.

Tao, F.; Esposito, A.; and Vinciarelli, A. 2020. Spotting the Traces of Depression in Read Speech: An Approach Based on Computational Paralinguistics and Social Signal Processing. *Proc. Interspeech 2020*, 1828–1832.

Tasnim, M.; and Stroulia, E. 2019. Detecting Depression from Voice. In *Canadian Conference on Artificial Intelligence*, 472–478. Springer.

Uher, R.; Cumby, J.; MacKenzie, L. E.; Morash-Conway, J.; Glover, J. M.; et al. 2014. A familial risk enriched cohort as a platform for testing early interventions to prevent severe mental illness. *BMC psychiatry*, 14(1): 344.

Valstar, M.; Gratch, J.; Schuller, B.; Ringeval, F.; Lalanne, D.; Torres Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; and Pantic, M. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proc. ACM workshop on Audio/visual emotion challenge*, 3–10.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Walker, J.; Burke, K.; Wanat, M.; Fisher, R.; Fielding, J.; et al. 2018. The prevalence of depression in general hospital inpatients: a systematic review and meta-analysis of interview-based studies. *Psychological medicine*, 48(14).

Wan, L.; Wang, Q.; Papir, A.; and Moreno, I. L. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4879–4883. IEEE.

Williamson, J. R.; Quatieri, T. F.; Helfer, B. S.; Ciccarelli, G.; and Mehta, D. D. 2014. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*.

Yang, L.; Jiang, D.; He, L.; Pei, E.; Oveneke, M. C.; and Sahli, H. 2016. Decision tree based depression classification from audio video and language information. In *Proc. ACM workshop on Audio/visual emotion challenge*, 89–96.