DepAudioNet: An Efficient Deep Model for Audio based Depression Classification

Xingchen Ma
IRIP Lab, Beihang University
Beijing 100191, China

Hongyu Yang
IRIP Lab, Beihang University
Beijing 100191, China

Qiang Chen
IRIP Lab, Beihang University
Beijing 100191, China

Di Huang*
IRIP Lab, Beihang University
Beijing 100191, China

Yunhong Wang
IRIP Lab, Beihang University
Beijing 100191, China

ABSTRACT

This paper presents a novel and effective audio based method on depression classification. It focuses on two important issues, *i.e.* data representation and sample imbalance, which are not well addressed in literature. For the former one, in contrast to traditional shallow hand-crafted features, we propose a deep model, namely DepAudioNet, to encode the depression related characteristics in the vocal channel, combining Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to deliver a more comprehensive audio representation. For the latter one, we introduce a random sampling strategy in the model training phase to balance the positive and negative samples, which largely alleviates the bias caused by uneven sample distribution. Evaluations are carried out on the DAIC-WOZ dataset for the Depression Classification Sub-challenge (DCC) at the 2016 Audio-Visual Emotion Challenge (AVEC), and the experimental results achieved clearly demonstrate the effectiveness of the proposed approach.

CCS Concepts

•Computing methodologies → Activity recognition and understanding;

Keywords

Depression Recognition, Audio Representation, CNN, LSTM

1. INTRODUCTION

Major Depressive Disorder (MDD), usually simply named depression, is a mental disorder characterized by a pervasive and persistent low mood, accompanied by low self-esteem as well as a loss of interest or pleasure in normally enjoyable activities. It has negative impacts on a person's family, work or school life, sleeping and eating habits, and general health. When depression is left untreated, it can cause severe consequences, *e.g.* addiction, self-injury, reckless behavior, and even suicide. For its harmfulness, in recent years, MDD has received increasing attention in many related communities.

Fortunately, medical studies [29] [24] show that depression is curable, and early detection of depression is very important to control it at an initial stage. Traditional approaches of depression analysis are prevailingly dependent on the verbal reports of patients, the behaviors reported by relatives or friends, and the mental status examinations, such as the Scale for the Assessment of Negative Symptoms (SANS [2]), the Hamilton Rating Scale for Depression (HRSD [38]) and the Beck Depression Inventory (BDI-II [26]). They all make use of subjective ratings, and due to the lack of objective and quantitative measurements, their results tend to be inconsistent at different times or in various environment. Besides, they commonly require extensive human expertise and are time-consuming. Therefore, it becomes a necessity to focus on Automatic Depression Detection (ADD).

ADD is a young topic and has not been investigated until 2009 [8]. The following studies mainly analyze audio signals, video signals, or both of them (see Sec.2 for more details). In general, these methods employ a framework that first extracts affect related features from the vocal or/and visual channel, and then builds certain classifiers or regressors for prediction. However, to the best of our knowledge, there are two crucial issues that are not well addressed in literature, limiting the ADD performance. On the one hand, the majority of the previous attempts represent the vocal or visual properties by using hand-crafted features. For instance, the spectrum, energy, and Mel-Frequency Cepstrum Coefficients (MFCCs) are widely exploited to encode the cues in the audio modality, while Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) are the representative ones to describe the facial characteristics in the video modality. They are designed based on the knowledge of human beings in the specific domain, and report promising results. Nevertheless, since our cognition of depression is still not so adequate, such features probably incur partial representation of audio or video data, which leaves space for improvement. More recently, deep learning techniques have demonstrated their competency at many audio and video based applications. They hierarchically learn high-level abstracted features through a deep graph with multiple processing layers and prove substantially superior to the shallow feature based ones. Some attempts [16, 7] have been made

*indicates the corresponding author (dhuang@buaa.edu.cn).

AVEC'16, October 16, 2016, Amsterdam, The Netherlands.

DOI: http://dx.doi.org/10.1145/2988257.2988267

for continuous affect recognition of two dimensions: Arousal and Valence, but little progress is found in depression prediction. On the other hand, most of current benchmarks suffer from data imbalance between positive (*i.e.* depression) and negative samples (as in AVEC 2016, the DCC challenge) or among different depression intensities (as in AVEC 2013, the DSC challenge), leading to a large bias in the classification or regression model. Furthermore, a much longer signal of an individual may highlight some person dependent properties that are not related to depression at all, making the circumstance even worse.

This paper presents a novel and effective approach to the Depression Classification Sub-Challenge (DCC) at the 2016 Audio-Visual Emotion Challenge (AVEC), aiming to classify whether a person is labeled as depressed or not using audio data. It encodes the depression cues of the vocal signal in a deep model, namely DepAudioNet, consisting of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). The one-dimensional convolutional layer in CNN incorporates short-term temporal and spectral correlations, the one-dimensional max-pooling layer captures middle-term correlations and LSTM extracts long-term correlations, producing a more comprehensive audio representation. Meanwhile, we introduce a random sampling strategy to train the deep vocal model, which largely alleviates the bias caused by the imbalance in uneven sample distribution. Experiments are conducted on the DAIC-WOZ dataset used for the AVEC 2016 competition, and the performance is greatly superior to the counterpart in the baseline, illustrating the effectiveness of the proposed approach.

2. RELATED WORK

Although there exist a large number of studies that dedicate to model the correlation between the emotional states and the properties of visual or vocal clues, the effort on depression classification (or depression prediction) is not that extensive. In this section, we briefly review the recent methods on ADD, which can be roughly categorized into audio based, video based, and multi-modal based, according to the information adopted.

Video based methods employ spatial and temporal information in the visual channel, where dynamic features are extracted from videos to capture depression related facial and body motions, which are further fed into a standard classification and regression backend to predict the depression level. As far as we know, the earliest published attempt [35] on automatic emotion modeling is based on the visual modality. The authors aimed at diagnosing schizophrenia, another mental disorder more serious than depression, and proposed a computational framework that creates probabilistic expression profiles for video data and potentially helps to automatically quantify emotional differences between patients with neuropsychiatric disorders and healthy controls. Geometric features were extracted based on the facial landmarks and several probabilistic classifiers were trained. To incorporate temporal information, they propagated classification results at each frame throughout the whole video, claiming that temporal dynamics are essential to capture subtle changes of facial expressions. Regarding depression itself, the first study [8] follows this framework, where manual Facial Action Coding System (FACS) and Active Appearance Model (AAM) were adopted to represent facial expressions, and Support Vector Machine (SVM) and logistic regression were exploited for decision making respectively. Its finding suggests the feasibility of ADD and has positive impacts on the clinical theory and practice. Recently, Cummins *et al.* [9] compared Space-Time Interest Points (STIPs) and Pyramid of Histogram of Gradients (PHOG) in their Support Vector Regression (SVR) based depression prediction system and found PHOG performed better in capturing visual variations. Wen *et al.* [36] extracted Local Phase Quantization at Three Orthogonal Planes (LPQ-TOP) for representing dynamic clues and utilized sparse coding and SVR to the predict depression level. They both reported very competitive results on the DSC challenge at AVEC 2013.

Compared with the features in the visual modality that convey the temporal information, which is crucial for modeling the facial expression changes, variations of speech production, such as rhythm, stress, and intonation, also indicate the changes of the emotional state and mental condition. An early study shows that noticeable acoustic changes could be aroused by even slight physiological and cognitive changes [32]. Therefore audio-based ADD methods [37], [8], [9], [21] are investigated to distinguish the arousal state of the speaker. They analyze the audio features, such as prosodic and acoustic features, formant features, spectral, *etc.*, which are supposed to be related to the depression emotion. Williamson *et al.* [37] explored two vocal tract representations, *i.e.* formant-frequency tracks and Mel-cepstral features, to encode the vocal tract resonant frequencies and spectral shape dynamics. With the two feature sets reflecting the changes in coordination of vocal tract motion associated with MDD, a Gaussian mixture model (GMM)-based multivariate regression scheme was then designed to make the final prediction. Later, they enhanced the approach by generating high-level features from the low-level ones by a multi-scale correlation structure and timing feature sets. In [33], Scherer *et al.* investigated four voice quality features as biomarkers for psychological distress, which are discriminative on a breathy to tense dimension, and used an SVM for classification. Besides, several studies were dedicated to the performance of diverse acoustic features in terms of depression prediction. In [28], a wide range of vocal features are explored, including estimated articulatory trajectories, acoustic characteristics, *etc.*, combined with back-ends of SVR, Gaussian, and decision trees, and comparable depression scale rating results were achieved on the AVEC-2014 development set by using the vocal channel only. For a more thorough review on ADD using the audio modality, refer to [10].

The improvements of the audio based methods and the vision based ones in depressive disorder analysis are always staggered, and the recent years have witnessed the achievements in fusing these two types, and several studies have suggested that such combination of the two complementary modalities improves the performance of depression detection. Recently, Meng *et al.* [27] applied Local Binary Patterns (LBP) [18] and Edge Orientation Histograms (EOH) to encode the dynamic visual cues represented in the Motion History Histogram (MHH) feature space, and then used Partial Least Square (PLS) to predict depression levels; whereas for audio processing, a set of spectral Low-Level Descriptors (LLD) features are explored to encode the characteristics of the audio, which are further fed to MHH to extract change information of the vocal expression, followed by the PLS based repression as does in the video process. The fusion results outperformed either modality, which ranked

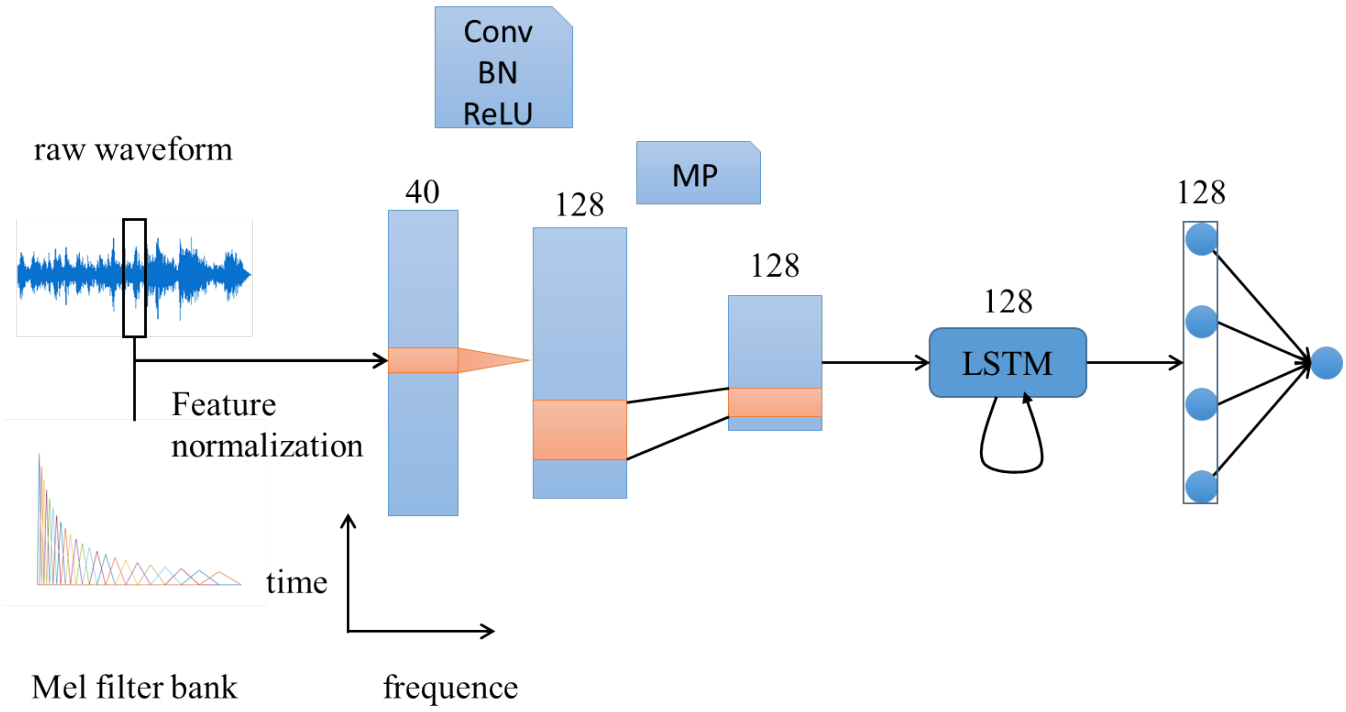

Figure 1: Flowchart of DepAudioNet.

the first place in video-based methods in the AVEC-2013 challenge. Kächele *et al.* [20] presented a hierarchical classifier framework, which stacked a multilayer neural network over the SVR ensemble, to recognize the depression state and adopted the Kalman filter for the final audio-video decision fusion, which improved the prediction accuracy on the AVEC-2013 dataset. Chao *et al.* [6], investigated the recent dominant deep learning models on this issue. The extracted audio video features were firstly fused in feature level as representation of the abnormal behavior, and then the Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) was exploited to describe dynamic temporal information. They used multi-task learning to boost the performance and reported competitive results on the AVEC-2014 dataset. This study indicated the promising future of exploiting the temporal information and both modalities in automatic depression detection.

3. DEEP AUDIO DEPRESSION MODEL

Considering the superiority of deep features over the traditional shallow ones, this paper presents a novel deep model, namely DepAudioNet, to encode the depression related temporal clues in the vocal modality and predict the presence of depression of a test audio sample. The entire framework, key components (*i.e.* CNN and LSTM), and random sampling training strategy, are described in the subsequent.

3.1 Approach Overview

The framework of DepAudioNet is illustrated in Figure 1, and DepAudioNet is a serial combination of CNN and LSTM, and it is thus expected to not only produce a high-level representation of the raw waveform, but also to capture short-term and long-term temporal variability.

Pre-processing is applied to the audio samples in advance. In each audio file, the long-lasting pauses are removed according to the timestamps located by a silence detector and the rest slices containing speech of the individual are linked together to generate a single file.

The Mel-scale filter bank feature, a low level audio descriptor, is employed to preliminarily represent the vocal signal. Each audio signal is split into several non-overlapping segments, on each of which a number of Mel-scale filter bank features are computed. After standard normal variate normalization, all the responses from the same segment are concatenated along the time axis, constructing a time-frequency $2D$ representation (as shown in Figure 2-(b)(d), which is further fed into DepAudioNet as input.

In DepAudioNet, a one-dimensional convolution layer is first exploited in the network structure, whose kernel size k is generally smaller than 3, suggesting that a number of short-term features are captured at this layer. Batch Normalization (BN) [19] is then performed to make the intermediate presentations subject to a standard normal distribution, accompanied with the reduced internal covariate shift [19] and regularized network. The rectified activation further introduces a nonlinear transformation and sparsity, followed by a one-dimensional max-pooling layer and a dropout layer. The pooling operation not only provides small translation invariance on the time axis, and more importantly, it neatly handles the middle-term temporal correlations by replacing the value at a certain location with a summary statistic of the outputs along a longer time window. An LSTM layer and two fully connected layers are stacked at the end of the network architecture, for the purpose of encoding long-range variability along the time axis and making the prediction. Combined with the convolution and max-pooling operation,

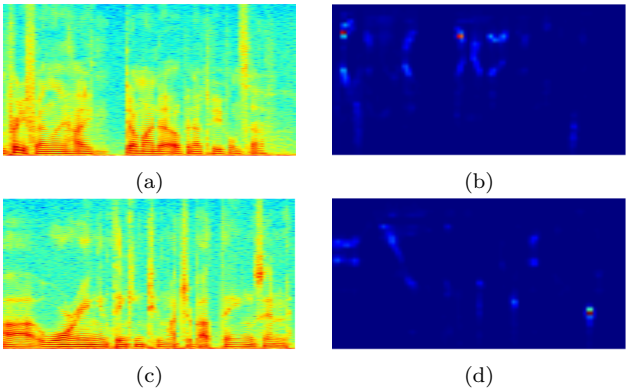

Figure 2: Spectrogram and Mel-scale filter bank feature visualization. (a) and (b) are spectrogram and filter bank feature of one audio slice from the non-depressed class respectively, (c) and (d) are spectrogram and filter bank feature of one audio slice from the depressed class.

it provides a hierarchical representation that comprehensively models the temporal properties in the vocal modality. The cost function is binary cross-entropy, and the Stochastic Gradient Descent (SGD) algorithm is used for optimization. By a majority vote method over the whole segments, the depression prediction of an audio is finally made.

3.2 Convolutional Neural Networks

CNNs have achieved breakthroughs [22] in image recognition recently and have been widely used as a powerful tool for many related tasks, such as speech recognition and signal processing [4].

Generally, a typical CNN contains one or more pairs of convolution and max-pooling layers. A convolution layer's parameters consist of a set of learnable filters, which have a small receptive field replicated along the whole input space. A max-pooling layer partitions convolution layer activations into non-overlapping rectangles and takes the maximum filter activation from these sub-regions.

There are two significant ideas making the convolution layer useful and effective, *i.e.* local connectivity and weight sharing. Local connectivity restricts that each neuron connects to only a local region of inputs, leading to sparse interactions [3]; and weight sharing reduces the number of the parameters and makes CNN much more efficient than regular feedforward multilayer perceptrons. These two ideas are illustrated in Fig. 3, where each node merely receives inputs from the three nearest or local inputs, and moreover, the weights w_1 to w_3 are restricted to be identical on the same output feature map. The output o_j can be expressed:

$$o_j = f(\sum_{i=1}^{3} w_i x_{i+j-1}) \tag{1}$$

where w_i stands for the weight, and x_{i+j-1} is the ith input of node j, and $f(\cdot)$ is a non-linear activation function like the sigmoid function or Rectified Linear Unit (ReLU).

It is common to deploy a pooling layer after a convolution one, replacing the output of the net at a certain location with a summary statistic of the nearby outputs [3]. The pooling layer aims to reduce the resolution of feature maps and introduce invariance to small variations in location. Max-pooling

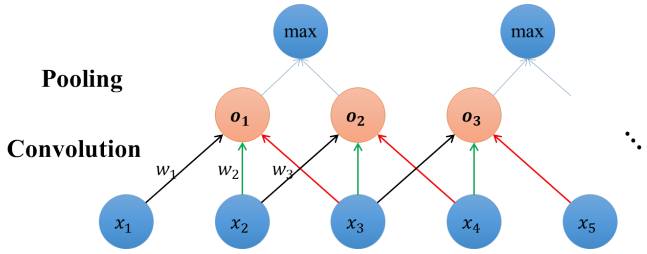

Figure 3: Illustration of local connectivity and weight sharing in convolution network.

is one of the most popular pooling operations, as in the upper part of Fig. 3, where each node is calculated by taking the max value on the corresponding region.

Pooling is a form of non-linear down-sampling, and it provides the translation invariance and tolerance to minor differences of positions of object parts, which is quite essential in this specific case of ADD, due to the fact that we care more about whether depression feature is present than where it is exactly in the audio signal [3].

There exist a number of studies [31, 13, 5, 30] that directly learn acoustic deep models from raw waveforms as the manner in image recognition where raw pixels are usually the inputs to CNN. But a recent finding [31] shows it is more efficient to build the models using some low-level audio descriptors. MFCCs is one of the most popular audio representations, whereas it is not suitable in this case, since Discrete Cosine Transform (DCT) projects the spectral energies into a new basis, which may not maintain the locality [1] required by CNN. Based on such observations, we exploit the Mel-scale filter bank features as the input. Compared to MFCCs, Mel-scale filter bank is much more appropriate for local filtering in this CNN configuration. The Mel-scale filter bank feature is computed by multiplying Short-Time Fourier Transform (STFT) magnitude coefficients with the corresponding filter, and it can thus be regraded as a non-linear transformation of spectrogram. It builds 40 log-spaced filters according to the following Mel-scale:

$$Mel(freq) = 2595 \cdot \log_{10}(1 + \frac{freq}{700}) \tag{2}$$

The filter bank and spectrogram features of two audio slices in the DAIC-WOZ database are visualized in Figure 2-(b)(d) and (a)(c) respectively.

In the traditional CNNs for images recognition, the interleaving convolution and pooling layers are used, and convolution kernel size and pooling size are often square. For the audio signal, which can also be represented as an 2D spectrograms over time and frequency, things are different. In spectrogram representation, time axis and frequency axis have different meanings, and they are asymmetric compared to width and height in regular RGB or gray images. The same spectral patterns in different frequency bands could indicate totally distinct audio classes, and square convolution and pooling in CNN used for image recognition would cause confusion among different audio classes and weaken the discriminative ability [11]. Therefore, in this study, we attempt to use the one-dimensional convolution along the whole frequency axis instead of square size filter to overcome this problem. In particular, the convolution filter size is $40 \times k$

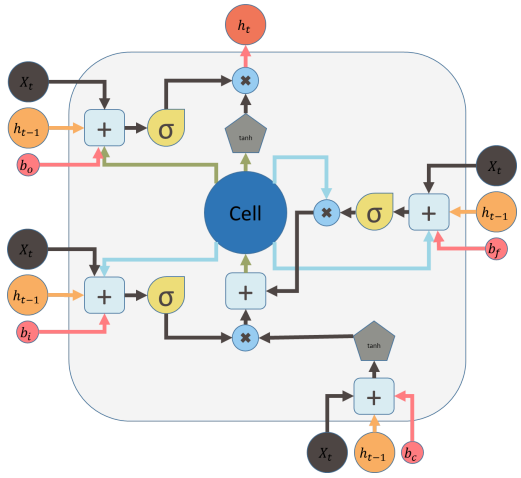

Figure 4: Structure of LSTM Cell [1].

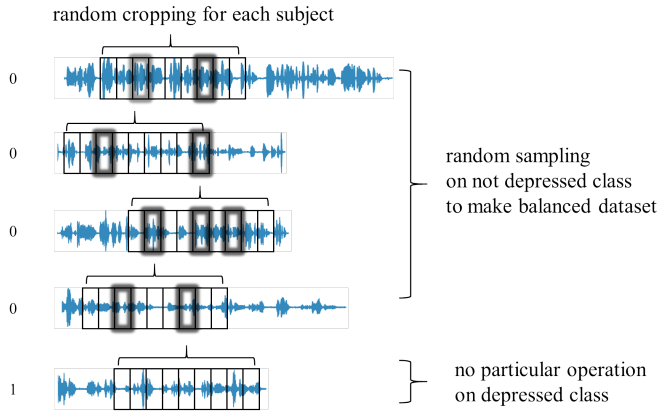

Figure 5: Generation procedure of a mini-batch in the training phase (the depressed and not depressed classes are denoted by "1" and "0" respectively).

for Mel-scale filter bank, where k is the convolution kernel width along the time axis associated with applications. And accordingly, we use one-dimensional max-pooling. For audio signal processing, one dimensional convolution handles the short-term temporal correlations, and one-dimensional max-pooling not only adds invariability along the temporal axis, but also covers middle-term temporal correlations.

3.3 Long Short-Term Memory

In the previous studies as introduced in Sec.2, when the features of the audio signal are available, SVM, tree-based classifier, and logistic regression are often introduced for depression prediction and report state of the art results. However, they do not explicitly consider time variability. They generally assume the independence of samples and are sequence-agnostic [25]. For sequence samples, such as audio signals, where a long-range dependency exists, these methods tend to lose useful information.

Hidden Markov Model (HMM), a statistical Markov model where an observed sequence is modeled with unobserved states, is capable of capturing time dependencies and has achieved great success in speech analysis. Due to the Markov assumption, each hidden state depends only on the immediately previous state, and thus makes it difficult to model long-range dependencies in sequence data. Although HMM can take a larger context window into account, it will make the state space exponentially grow with the size of the window [25]. Besides, the time complexity of Viterbi, a dynamic programming algorithm to perform efficient inference in HMM, is $O(T \times |S|^2)$, where T is the sequence length and S is the state space. When the state space is large, HMM becomes infeasible.

Recurrent Neural Networks (RNN) do not make the Markov assumption in theory, and they can capture long-term dependencies. LSTM [17] is adopted to overcome the vanishing gradients problem, and it has the ability to model much longer temporal structure than a vanilla RNN by introducing memory cells. The LSTM cell used in this paper is based on [15], illustrated in Fig. 4. The computation in the LSTM model proceeds according to:

[1]This picture is credited to https://github.com/shi-yan/FreeWill/tree/master/Docs/Diagrams

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (6)$$

$$h_t = o_t \odot tanh(c_t) \quad (7)$$

where σ is the sigmoid function, $\odot$ is element-wise multiplication, W is the weight matrix whose subscripts have obvious meaning, i, f, o and c are the input gate, forget gate, output gate, and cell memory respectively, and b_i, b_f, b_o, and b_c are corresponding bias terms.

The LSTM layer used in this paper is just stacked after the final CNN layer and then follows two full connected layers, which is illustrated at the end of Figure 1.

3.4 Random Sampling

A major problem in learning shallow or deep depression models lies in uneven sample distribution. Many current benchmarks suffer from data imbalance between positive and negative samples or among different depression levels, which incurs a large bias in classification or regression. For example, in the DAIC-WOZ database provided by AVEC 2016 for the DCC challenge, the number of non-depressed subjects is about four times bigger than that of depressed ones in both training and development parts. If these samples are directly adopted for learning, the model will have a strong bias to the non-depressed class, thus making it not reliable. Moreover, regarding the length of each sample, a much longer signal of an individual may emphasize some characteristics that are person specific, which tends to deteriorate the situation.

To solve this problem, we introduce a simple yet effective scheme by random sampling for training as in Fig. 5.

In our case, we first conduct random cropping on each pre-processed audio file, and make the inputs to CNN have equal proportion for every subject. This operation is to minimize the influence of characteristics of individual subjects. We then split the randomly cropped slices into equal parts.

Table 1: Performance of Mel-scale filter bank features of different parameters (values of non-depression in brackets).

Partition	time window W	max-pooling length l	F1 score	Precision	Recall
Baseline [34]	-	-	0.41 (0.58)	0.27 (0.94)	0.89 (0.42)
Development	100	3	0.39(0.64)	0.26(0.88)	0.72(0.50)
Development	120	3	**0.52(0.70)**	0.35(1.00)	1.00(0.54)
Development	140	3	0.36(0.71)	0.27(0.85)	0.57(0.61)
Development	100	5	0.38(0.73)	0.29(0.86)	0.57(0.64)
Development	120	5	0.42(0.70)	0.29(0.89)	0.71(0.57)
Development	140	5	0.35(0.68)	0.25(0.84)	0.57(0.57)

Table 2: Performance of the spectrogram features of different parameters (values of not-depression in brackets).

Partition	time window W	max-pooling length l	F1 score	Precision	Recall
Baseline [34]	-	-	0.41 (0.58)	0.27 (0.94)	0.89 (0.42)
Development	100	3	0.38(0.64)	0.26(0.88)	0.71(0.50)
Development	120	3	0.48(0.63)	0.31(1.00)	1.00(0.46)
Development	140	3	0.33(0.77)	0.27(0.83)	0.43(0.71)
Development	100	5	0.30(0.72)	0.23(0.82)	0.43(0.64)
Development	120	5	**0.52(0.70)**	0.35(1.00)	1.00(0.54)
Development	140	5	0.40(0.76)	0.31(0.86)	0.57(0.68)

For non-depressed samples, we perform random sampling to make the size match the one of depressed samples. For Mel-scale filter bank features, every rectangle produces a $40 \times W$ representation in our configuration, this representation becomes one sample in mini-batch. Referring to Fig. 5, every mini-batch contains 18 samples, and half of them come from the depressed class while the other half from the non-depressed class. Finally, we shuffle every mini-batch.

Random sampling contributes to dealing with data imbalance in two-fold. For the first, it minimizes the effects of individual subjects by randomly cropping on the audio signal to guarantee that the inputs to the network have equal parts from every subject. For the second, it randomly picks samples from non-depressed classes to compose mini-batches which have equal parts from two classes. These two operations make the network generalize well.

4. EXPERIMENTAL EVALUATION

To evaluate the effectiveness the proposed DepAudioNet, we carry out extensive experiments on the DAIC-WOZ dataset, in the Depression Sub-Challenge at AVEC 2016. The dataset, experimental protocols, and prediction results are introduced in the following subsections.

4.1 Dataset

DAIC-WOZ is part of a large corpus, namely the Distress Analysis Interview Corpus (DAIC) [14], which contains clinical interviews designed to support the diagnosis of psychological distress conditions, such as anxiety, depression and post-traumatic stress disorder. The interviews are collected by a computer agent that interacts with people and identifies verbal and non-verbal indicators of mental illness [12]. This collection includes audio and video recordings and extensive questionnaire responses, where the part of the corpus contains the Wizard-of-Oz interviews, conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room. Samples are transcribed and annotated for a variety of verbal and non-verbal features.

4.2 Parameter Setting

The Depression Sub-Challenge (DSC) at AVEC 2016 attempts to detect if depression exists in the human-computer interaction as indicated by the PHQ-8 score [23], which is a self-reported heath scale. It is a binary classification task, and the the averaging *F1 score* is used for assessment.

In feature extraction, 40 Mel-scale filters are calculated on each audio segment in total. In particular, the frequency ranges from 130 Hz to 6854 Hz, the size of the Hanning window W_{fft} is 1024, and the audio sample rate is 16000 Hz, which leads to a covering domain of 1024/16000 Hz=0.064s; accordingly, the hop size is 32ms, half of the analysis window. The time window size of the segment W is 120, which is determined in cross-validation on the development set, and the 120 normalized responses from the same segment are concatenated along the time axis, covering an audio clip of $(W \times W_{fft}/2 + 1)/16000 \approx 3.84$s. Meanwhile, besides the Mel-scale bank filter features, we also adopt another transformation of the spectrogram, which is the magnitude of STFT of audio signal, to evaluate the impacts of LLD features, and the same configuration is employed as for the Mel-scale filter bank features, except that the feature dimension is 513.

In DepAudioNet, we apply 128 convolution maps being of

3×1 on the time-frequency of 2D representation capturing the variability of 64ms precision, with the max-pooling operation on a region being size of 3×1, incorporating longer time correlations, and the node numbers in the last two fully connected layers are set to 128 and 1, respectively. In the step of random sampling, each mini-batch contains 18 vocal samples, which come evenly from both classes.

4.3 Results

We analyze the experimental results at three layers: (1) comparison with the baseline; (2) parameter analysis; and (3) comparison of diverse low level audio features.

Table 1 displays an overview of the depression classification results based on the Mel-scale filter bank features with varying parameter values. From this table we can see that the best performance in terms of *F1 score* achieved by the proposed DepAudioNet is 0.52 (0.70) with the time window and max-pooling length set at 120 and 3 respectively, and it is obviously superior to the averaging baseline 0.41 (0.58). Besides, the corresponding precision and recall also outperform the ones in the baseline. Particularly, the precision for class *non-depressed* is up to 100%, which clearly proves the effectiveness of the proposed method.

Furthermore, it can be seen from Table 1 that with the max-pooling length fixed, the best results are always reached when the time window size W is set to 120. The reason lies in that the proposed DepAudioNet requires a proper compromise between the long-range time variety and the modeling capability. Theoretically, LSTM is able to handle the input of arbitrary length, whereas the performance declines when encountering rather long sequences in practice. Conversely, the audio clips do not convey much information about the variability along time axis if the window size is quite small, which weakens the model's ability in discrimination. While with the time window fixed, a smaller max-pooling length generally gives better results.

Finally, we conduct the same experiment using the spectrogram feature, and the results are presented in Table 2. The spectrogram feature reaches the top performance, with the max-pooling length set at 5 and the time window size at 120. When we compare their corresponding results in Table 1 and 2, we can see that the figures are comparable. Considering the fact that Mel-scale filter bank feature is a further non-linear transformation of spectrogram, DepAudioNet indeed learns such information that is beneficial to good audio representations.

5. CONCLUSION

In this paper, a novel deep neural network, DepAudioNet is proposed for ADD. This hierarchical structure delivers a comprehensive audio representation by capturing the short-term and middle-term temporal and spectral correlations with CNN, and extracting the long-term correlations via LSTM. Moreover, a random sampling strategy is adopted to balance the uneven sample distribution in this specific case. Evaluations are carried out on DAIC-WOZ used for the AVEC 2016 competition, and the results demonstrate the effectiveness of the proposed method.

6. ACKNOWLEDGMENT

This work was supported in part by the national key research and development plan under Grant 2016YFC0801002, the Hong Kong, Macao, and Taiwan Science and Technology Cooperation Program of China under Grant L2015TGA9004, and the National Natural Science Foundation of China under Grant 61540048.

7. REFERENCES

[1] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu. Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 22(10):1533–1545, Oct. 2014.

[2] N. C. Andreasen. Scale for the Assessment of Negative Symptoms (SANS). *The British Journal of Psychiatry*, 1989.

[3] I. G. Y. Bengio and A. Courville. Deep learning. Book in preparation for MIT Press, 2016.

[4] Y. Bengio. A connectionist approach to speech recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 07(04):647–667, Aug. 1993.

[5] M. Bhargava and R. Rose. Architectures for deep neural network based acoustic models defined over windowed speech waveforms. In *International Speech Communication Association*, 2015.

[6] L. Chao, J. Tao, M. Yang, and Y. Li. Multi task sequence learning for depression scale prediction from video. In *Affective Computing and Intelligent Interaction*, pages 526–531. IEEE, 2015.

[7] S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *International Workshop on Audio/Visual Emotion Challenge*, pages 49–56. ACM, 2015.

[8] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. La Torre. Detecting depression from facial actions and vocal prosody. In *International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[9] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps. Diagnosis of depression by behavioural signals: a multimodal approach. In *International Workshop on Audio/Visual Emotion Challenge*, pages 11–20. ACM, 2013.

[10] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, July 2015.

[11] L. Deng, O. Abdel-Hamid, and D. Yu. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6669–6673, May 2013.

[12] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. M. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. R. Traum, R. Wood, Y. Xu, A. A. Rizzo, and L. Morency. Simsensei kiosk: a virtual human interviewer for healthcare decision support. In *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France,*

May 5-9, 2014, pages 1061–1068, 2014.

[13] P. Golik, Z. Tüske, R. Schlüter, and H. Ney. Convolutional neural networks for acoustic modeling of raw time signal in LVCSR. In *Interspeech 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 26–30, 2015.

[14] J. Gratch, R. Artstein, G. Lucas, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L.-P. Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of Language Resources and Evaluation Conference*, pages 3123–3128, 2014.

[15] A. Graves. Generating sequences with recurrent neural networks. *arXiv:1308.0850 [cs]*, Aug. 2013.

[16] L. He, D. Jiang, and H. Sahli. Multimodal depression recognition with dynamic visual and audio cues. In *International Conference on Affective Computing and Intelligent Interaction*, pages 260–266, Sept. 2015.

[17] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. Local binary patterns and its application to facial image analysis: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 41(6):765–781, 2011.

[19] S. Ioffe and C. Szegedy. Batch Normalization: accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167 [cs]*, Feb. 2015.

[20] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. *depression*, 1(1), 2014.

[21] H. Kaya and A. A. Salah. Eyes whisper depression: a cca based multimodal approach. In *International Conference on Multimedia*, pages 961–964. ACM, 2014.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[23] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173, Apr. 2009.

[24] C. W. Lejuez, D. R. Hopko, and S. D. Hopko. A brief behavioral activation treatment for depression treatment manual. *Behavior Modification*, 25(2):255–286, 2001.

[25] Z. C. Lipton, J. Berkowitz, and C. Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv:1506.00019 [cs]*, May 2015.

[26] A. Mcpherson and C. R. Martin. A narrative review of the Beck Depression Inventory (BDI) and implications for its use in an alcohol-dependent population. *Journal of Psychiatric and Mental Health Nursing*, 17(1):19–30, Feb. 2010.

[27] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *International Workshop on Audio/Visual Emotion Challenge*, pages 21–30. ACM, 2013.

[28] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, and M. Graciarena. The SRI AVEC-2014 evaluation system. In *International Workshop on Audio/Visual Emotion Challenge*, pages 93–101. ACM, 2014.

[29] National Collaborating Centre for Mental Health (UK). *Depression: the treatment and management of depression in adults (updated edition)*. National institute for health and clinical excellence: guidance. British Psychological Society, 2010.

[30] D. Palaz, R. Collobert, and others. Analysis of cnn-based speech recognition system using raw speech as input. In *Interspeech*, 2015.

[31] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals. Learning the speech front-end with raw waveform cldnns. In *Interspeech 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1–5, 2015.

[32] K. R. Scherer. Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[33] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658, Oct. 2014.

[34] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proceedings of AVEC'16, co-located with ACM MM 2016*, Amsterdam, The Netherlands, October 2016. ACM.

[35] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma. Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of Neuroscience Methods*, 168(1):224–238, Feb. 2008.

[36] L. Wen, X. Li, G. Guo, and Y. Zhu. Automated depression diagnosis based on facial dynamic analysis and sparse coding. *IEEE Transactions on Information Forensics and Security*, 10(7):1432–1441, 2015.

[37] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *International Workshop on Audio/Visual Emotion Challenge*, pages 41–48. ACM, 2013.

[38] M. Zimmerman, I. Chelminski, and M. Posternak. A review of studies of the Hamilton Depression Rating Scale in healthy controls: implications for the definition of remission in treatment studies of depression. *The Journal of Nervous and Mental Disease*, 192(9):595–601, Sept. 2004.