Clinical Depression and Affect Recognition with EmoAudioNet

Emna Rejaibi∗, Daoud Kadoch†, Kamil Bentounes†, Romain Alfred‡, Mohamed Daoudi§, Abdenour Hadid¶, Alice Othmani∥
∗Institut National des Sciences Appliquées et de Technologie, Centre Urbain Nord BP 676-1080, Tunis, Tunisie
†Sorbonne Université, 75006 Paris, France.
‡cole Nationale Supérieure d’Informatique pour l’Industrie et l’Entreprise, ENSIIE, 91000, vry, France
§IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 CRIStAL, F-59000 Lille, France
¶Center for Machine Vision and Signal Analysis (CMVS) University of Oulu, Finland
∥Université Paris-Est, LISSI, UPEC, 94400 Vitry sur Seine, France

Corresponding author: alice.othmani@u-pec.fr

Abstract—Automatic analysis of emotions and affects from speech is an inherently challenging problem with a broad range of applications in Human-Computer Interaction (HCI), health informatics, assistive technologies and multimedia retrieval. Understanding human's specific and basic emotions and reacting accordingly can improve HCI. Besides, giving machines skills to understand human's emotions when interacting with other humans can help humans with a socio-affective intelligence. In this paper, we present a deep Neural Network-based architecture called EmoAudioNet which studies the time-frequency representation of the audio signal and the visual representation of its spectrum of frequencies. Two applications are performed using EmoAudioNet: automatic clinical depression recognition and continuous dimensional emotion recognition from speech. The extensive experiments showed that the proposed approach significantly outperforms the state-of-art approaches on RECOLA and DAIC-WOZ databases. The competitive results call for applying EmoAudioNet on others affects and emotions recognition from speech applications.

Index Terms—Socio-affective computing, Emotional Intelligence, speech emotion recognition, depression recognition, CNN, EmoAudioNet.

I. INTRODUCTION

Artificial Emotional Intelligence (EI) or affective computing has attracted increasing attention from the scientific community and hence automatic systems have been developed to recognize, interpret, process and simulate human affects. In fact, affect describes the experience of a human's emotion resulting from an interaction with stimuli. Humans express an affect through facial, vocal, or gestural behaviors. Giving machines skills of emotional intelligence is an important key to improve Human-Computer Interaction (HCI) or to understand human’s interactions with other humans. Thus, a wide range of applications has been developed in HCI, health informatics, assistive technologies and computer-vision applications.

Voice provides an elementary way by which humans or computers can detect and recognize emotions. Endowing computers with the ability to recognize vocal emotions can improve HCI. A happy, angry or fearful person will typically speak louder and faster, with strong frequencies, while a sad or bored person will speak slower with low frequencies. Emotional arousal and valence are the two main dimensional affects used to describe emotions. Valence describes the level of pleasantness and it is defined along a continuum from negative to positive, while arousal describes the intensity and the level of autonomic activation created by an event and it is defined along a continuum that varies from low (calm) to high (excited).

A final method for measuring a users affective state is to ask questions and to identify emotions during an interaction. Several post-interaction questionnaires exist for measuring affective states like the Patient Health Questionnaire (PHQ) for depression recognition and assessment. The PHQ is a multiple-choice self report questionnaire. It is a commonly used test composed of nine clinical questions. A score which range from 0 to 23 is assigned to describe Major Depressive Disorder (MDD) severity level.

Major Depressive Disorder is a mental disease which affects more than 300 million people in the world [1], i.e., 3% of the worldwide population. Each year, 800 000 suicides are enumerated [2] and a depressive state is the cause of 70% of them. Depression can affect people of all ages, genders, ethnicities and physical and social conditions. The psychiatric taxonomy classifies Major Depressive Disorder (MDD) among the low moods [3], i.e., a condition characterised by a tiredness and a global physical, intellectual, social and emotional slow-down. In this way, the depressive subjects have a tendency to have lower facial expressions. Their speech is slowed, the pauses between two speakings are lengthened and the tone of the voice (prosody) is more monotonous.

Because of its repercussions on the population and its impact on society, major depressive disorder is studied by diverse disciplines. Scientists from the artificial intelligence and computer vision communities have been paying increasing attention to develop computer-aided diagnostic tools for MDD. In this paper, a new deep neural networks architecture, called

EmoAudioNet...
EmoAudioNet, is proposed and evaluated for real-life depression and affect recognition from speech. The remainder of this article is organised as follows. Section II introduces related works in affect and depression recognition from speech. Section III describes the details of the overall proposed method. Section IV describes the entire experiments and the extensive experimental results. Finally, the conclusion and future work are presented in Section V.

II. RELATED WORK

Competitions and challenges events aiming to compare multi-modal data processing and machine learning methods for automatic audiovisual emotion analysis from real-life data, helped to popularize the development of applications for affect and depression recognition. Specifically, we can cite the six successive editions, since 2013, of the Audio/Visual Emotion Challenge (AVEC) [15] and of the Emotion Recognition in the Wild challenge (EmotiW) [30]. Many databases of interviews have been employed in the different editions of AVEC, EmotiW and Affect-in-the-wild (Aff-Wild) challenges. Certain databases comprises videos recorded in a controlled setting (RECOLA, DAIC-WOZ, the AVEC databases) [10], [11], [22], [26], [31]. Others are recorded “in the wild”, i.e., not “captured in highly controlled recording conditions” [40], [32], such as the SEWA [23], [39], the Acted Facial Expressions in Wild (AFEW) [33] and Aff-Wild databases [40]. Different features are exploited and a profusion of classifiers have been experimented with, as well as different architectures of neural networks from audio signals. These methods can be generally categorized into two groups: hand-crafted features-based approaches and deep learning-based approaches.

A. Handcrafted features-based approaches

In this family of approaches, there are two main steps: feature extraction and classification. An overview of handcrafted features-based approaches for affect and depression assessment from speech is presented in Table I.

1) Handcrafted features: From the video interviews, the soundtrack is retrieved. Open-source tools such as the CoVAREP(v1.3.2) [10], [11], [43], the Octave toolbox [10], the Emotion and Affect Recognition (openEAR) [15], [23] and the YAAFE toolbox [21], [47] are next used for the extraction of the acoustic Low-Level Descriptors (LLD).

The extracted LLD are grouped into four main categories: the spectral, the cepstral, the prosodic, and the voice quality LLD. Several spectral features are extracted from speech for affect and depression assessment such as the Harmonic Model and Phase Distortion Mean (HMPDM0-24), the Deviations (HMPDD0-12), the energy slope, the Hammarberg index, spectral flux [31], the spectral slope, the spectral variation, the spectral decrease, the spectral flatness, the perceptual spread, the amplitude modulation and the Line Spectral Pairs (LSP) frequency [21], [23]. The most used cepstral features are the Mel-Frequency Cepstral Coefficients (MFCC) [9], [10], [22], [23], [31] along with their deltas and delta-deltas and the Linear Prediction Cepstral Coefficients (LPCC) [21]. The prosodic LLD play a significant role in identifying emotions and depression as they characterise the tone of the voice. These LLD are obtained through the dissociation of the spectral features such as the fundamental frequency (F0) [10], [31], the F0 envelope [23], the Formants (F1, F2, F3) [51], [34], [65], and the loudness of the speech [21], [23]. The last group of LLD contains voice quality features that describe the efficiency of the speech and the silent segments (noise). Most commonly extracted voice quality features are the Normalized Amplitude Quotient (NAQ), the Quasi Open Quotient (QQQ), the Harmonic difference H1-H2, the Harmonic difference H1-H3, the Parabolic Spectral Parameter (PSP), the Maxima Dispersion Quotient (MDQ) [10], the spectral tilt/slope of wavelet responses (peak-Slope), the shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd), the Jitter, and the Shimmer [23].

Based on the LLD, a set of statistical features are also calculated. For example, the maximum of certain LLD, their minimum, their variance, their standard deviation, their quartiles and the rise and fall times are considered in several works [11], [21]. The functionals are extracted from the whole soundtrack as well as from small segments. Low et al. [29] propose the experimentation of the Teager Energy Operator (TEO) based features. The TEO is defined in discrete form [37] as :

$$\psi(x(t)) = x'^2(t) - x(t)x''(t)$$  \hspace{1cm} (1)

where $x$ is a LLD, $x'$ its first derivative and $x''$ its second derivative, at a time $t$.

A comparison of the performances of the prosodic, spectral, glottal (voice quality), and TEO features for depression recognition is realized in [25] and demonstrates that the different features have similar accuracies. The fusion of the prosodic LLD and the glottal LLD based models seems to not significantly improve the results, or decreased them (-16,25 to +5,50% of the accuracy). However, the addition of the TEO features improves the performances to +31,35% (variation of +6,80 to +31,35% for depressive male and of -6,34 to +7,14% for depressive female). Furthermore, Bag-of-Text-Words (BoTW) are used to quantify the redundancy of words and Bag-of-Audio-Words (BoAW) to quantify audio signal patterns [11], [22]. The openBOX toolkit have been employed in the extraction of the bag-of-words. According to Eyben et al., the audio features can be summarised in a few features within the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [33], Valstar et al. [10], Ringeval et al. [11], [22] and Yang et al. [45] propose also the use of an extended version of the GeMAPS, called eGeMAPS.

2) Classification of Handcrafted features: Comparative analysis of the performances of several classifiers in depression assessment and prediction indicate that the use of an hybrid classifier using Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) model gave the best overall classification results [14], [25]. Different fusion methods, namely feature, score and decision fusion have been also investigated in [14] and it has been demonstrated that : first, amongst the
fuson methods, score fusion performed better when combined with GMM, HFS and MLP classifiers. Second, decision fusion worked best for SVM (both for raw data and GMM models) and finally, feature fusion exhibited weak performance compared to other fusion methods.

B. Deep learning-based approaches

Recently, approaches based on deep learning have been proposed to assess the level of depression using the Beck Depression Inventory II (BDI-II) [29], where only the MFCC are considered. In others approaches, raw audio signals are fed to deep neural networks [28]. An overview of deep learning-based methods for affect and depression assessment from speech is presented in Table II.

Several deep neural networks have been proposed: feed-forward neural network (FF-NN) [20], convolutional neural network (CNN) [46], deep convolutional neural network (DCNN) [16], deconvolutional neural network (DNN) [43], [50], long short-term memory recurrent neural network (LSTM-RNN) [20], [22], [29], [42], [46], bidirectional long short-term memory recurrent neural network (BLSTM-RNN) [20], [22], [44], deep bidirectional long short-term memory recurrent neural network (DBLSTM-RNN) [21], LSTM-CNN [28], BLSTM-CNN [44], DCNN-DNN [45], BLSTM with Attention mechanism [46], deconvolutional neural network multiple instance learning (DNN-MIL) [44], BLSTM-MIL [44] and Capsule Network [46].

A comparative study [44] of some neural networks, BLSTM-MIL, BLSTM-RNN, BLSTM-CNN, CNN, DNN-MIL and DNN, demonstrates that the BLSTM-MIL outperforms the other studied architectures. In Jain [46], the use of a BLSTM with Attention mechanism is proposed. In parallel, a Capsule Network is experimented. Following this method, the Capsule Network is the most efficient architecture, compared to the BLSTM with Attention mechanism and also to a CNN and a LSTM-RNN. For the assessment of the level of depression using the Beck Depression Inventory II (BDI-II), Yang et al. [16] exerts a DCNN using audio features. To the best of our knowledge, their approach outperforms all the existing approaches on DAIC-WOZ dataset with a Root Mean Squared Error (RMSE) of 1,802 and 4,590 for depressed male and depressed female respectively.

III. PROPOSED METHOD

A. Method overview

A variety of signals features can be extracted for audio classification and speech recognition either to represent it in the time domain or in frequency domain. Hence, features of both, frequency domain and time domain are jointly required for sound classification. In fact, the extraction of good parametric representation of a signal is important to produce a better recognition performance.

Short-time spectral analysis is the most common way to characterize the speech signal using the Mel-frequency Cepstral Coefficients (MFCC). MFCCs present the variation of the human ear’s critical bandwidths with frequency and they can capture the phonetically important characteristics of the speech by expressing the signal in the Mel frequency scale. This parametric description of the spectral envelope, represents the speech amplitude spectrum in a compact form and combines the advantages of the cepstrum analysis with a perceptual frequency scale based on critical bands.

Besides the acoustic motivation of the MFCC features, audio signals in their time-frequency representations, often present interesting patterns in the visual domain [51]. The visual representation of the spectrum of frequencies of a signal using its spectrogram, called voiceprints, shows a set of specific repetitive patterns. Spectrograms are 2D images representing sequences of spectra with time along one axis, frequency along the other, and brightness or color representing the strength of a frequency component at each time frame. It can be defined as an intensity plot (usually on a log scale) of the Short-Time Fourier Transform (STFT) magnitude. Therefore, three motivations are behind the proposed approach:

- short-time spectral analysis of the audio signal,
- audio signals classification in the visual domain,
- combining information from time domain, frequency domain and visual domain.

To answer those motivations in a more robust way, low-level features with high-level features are considered and three deep convolutional neural networks are proposed as shown in Fig. 1:

- An MFCC-based CNN for emotion recognition from speech.
- An audio spectrogram-based CNN for emotion recognition from speech,
- An MCFF and spectrogram aggregated deep neural networks for emotion recognition from speech.

EmoAudioNet is designed for automatic emotion recognition and classification from speech. In this work, it is only evaluated on automatic clinical depression recognition and assessment and on spontaneous and continuous emotion recognition from speech. In the following, more details about the three architectures are given.

1) Pre-processing: a 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.

2) Spectrogram-based CNN architecture: The spectrogram-based CNN aims to classify audio features in the visual domain. The texture-like time-frequency representation usually contains distinctive patterns that capture different characteristics. The spectrogram-based CNN presents low-level features descriptor followed by a high-level features descriptor.

Low-level features descriptor. The spectrogram of an audio
**TABLE I: Overview of Shallow Learning based methods for Affect and Depression Assessment from Speech.**

| Ref | Features | Classification | Dataset | Metrics | Value |
|-----|----------|----------------|---------|---------|-------|
| Jan et al. [9] | spectral LLD + MFCC | PCA + PLS regression | AVEC2014 | RMSE MAE | 8.08 6.52 |
| Valstar et al. [10] | prosodic LLD + voice quality LLD + spectral LLD | SVM + grid search + random forest | DAIC-WOZ | F1-score Precision Recall | 0.410 (0.582) 0.267 (0.941) 0.889 (0.421) 7.78 5.72 |
| Dhall et al. [21] | energy LLD + spectral LLD + voicing quality LLD + duration features | non-linear chi-square kernel | AFEW 5.0 | unavailable | unavailable |
| Ringeval et al. [11] | prosodic LLD + voice quality LLD + spectral LLD | random forest | SEWA | RMSE MAE | 7.78 5.72 |
| Pampouchidou et al. [12] | prosodic LLD + voice quality LLD + spectral LLD | PCA + linear discriminant analysis | AVEC2013 + AVEC2014 | F1-score Precision Recall | 0.641 0.600 0.688 |
| Haq et al. [24] | energy LLD + prosodic LLD + spectral LLD + duration features | Sequential Forward Selection + Sequential Backward Selection + linear discriminant analysis + Gaussian classifier uses Bayes decision theory | Natural speech databases | Accuracy | 66.5% |
| Jiang et al. [13] | MFCC + prosodic LLD + spectral LLD + glottal features | ensemble logistic regression model for detecting depression E algorithm | hand-crafted dataset | Males accuracy Males sensitivity Males specificity Females accuracy Females sensitivity Females specificity | 81.82% 78.13% 85.29% 79.25% 70.19% 70.59% |
| Low et al. [25] | teager energy operator based features | Gaussian mixture model + SVM | hand-crafted dataset | Males accuracy Males sensitivity Males specificity Females accuracy Females sensitivity Females specificity | 86.64% 80.83% 92.45% 80.64% 78.87% 77.27% |
| Alghowinem et al. [14] | energy LLD + formants + glottal features + intensity LLD + MFCC + prosodic LLD + spectral LLD + voice quality LLD | Gaussian mixture model + SVM + decision fusion | hand-crafted dataset | Accuracy | 91.67% |
| Valstar et al. [15] | duration features + energy LLD + local minima/maxima related functionals + spectral LLD + voicing quality LLD | correlation based feature selection + SVR + 5-fold cross-validation loop | A ViD-Corpus | RMSE MAE | 14.12 10.35 |
| Valstar et al. [26] | duration features + energy LLD + local minima/maxima related functionals + spectral LLD + voicing quality LLD | SVR | AVEC2014 | RMSE MAE | 11.521 8.934 |
| Liu et al. [17] | energy LLD + MFCC + prosodic LLD + spectral LLD | filter thanks the minimal-redundancy-maximal relevance criterion + Sequential Forward Floating Selection + k-NN + SVM | hand-crafted dataset | unavailable | unavailable |
| Cummins et al. [18] | MFCC + prosodic LLD + spectral centroid | SVM | AVEC2013 | Accuracy | 82% |
| Lopez Otero et al. [19] | energy LLD + MFCC + prosodic LLD + spectral LLD | SVR | AV DLC | RMSE MAE | 8.88 7.02 |
| Meng et al. [27] | spectral LLD + energy LLD + MFCC + functionals features + duration features | PLS regression | AVEC2013 | RMSE MAE CORR | 11.54 9.78 0.42 |
| Yang et al. [29] | prosodic LLD + voice quality LLD duration features | Min-Redundancy Max-Relevance algorithm | hand-crafted dataset | Accuracy | 44.44% |
TABLE II: Overview of Deep Learning based methods for Affect and Depression Assessment from Speech.

| Ref               | Features                                                                 | Classification          | Dataset    | Metrics               | Value               |
|-------------------|--------------------------------------------------------------------------|-------------------------|------------|-----------------------|---------------------|
| Yang et al. [16]  | spectral LLD + cepstral LLD + prosodic LLD + voice quality LLD + statistical functionals + regression functionals | DCNN                    | DAIC-WOZ   | Depressed female RMSE | 4.590               |
|                   |                                                                          |                         |            | Depressed female MAE  | 3.589               |
|                   |                                                                          |                         |            | Not depressed female RMSE | 2.864               |
|                   |                                                                          |                         |            | Not depressed female MAE | 2.393               |
|                   |                                                                          |                         |            | Depressed male RMSE   | 1.802               |
|                   |                                                                          |                         |            | Depressed male MAE    | 1.690               |
|                   |                                                                          |                         |            | Not depressed male RMSE | 2.827               |
|                   |                                                                          |                         |            | Not depressed male MAE | 2.575               |
| Al Hanai et al. [42] | spectral LLD + cepstral LLD + prosodic LLD + voice quality LLD + functions | LSTM-RNN                | DAIC       | F1-score              | 0.67                |
|                   |                                                                          |                         |            | Precision             | 1.00                |
|                   |                                                                          |                         |            | Recall                | 0.50                |
|                   |                                                                          |                         |            | RMSE                  | 10.03               |
|                   |                                                                          |                         |            | MAE                   | 7.60                |
| Dham et al. [43]  | prosodic LLD + voice quality LLD + functionals + BoTW                    | FF-NN                   | AVEC2016   | RMSE                  | 7.631               |
|                   |                                                                          |                         |            | MAE                   | 6.2766              |
| Salekin et al. [44] | spectral LLD + MFCC + functionals                                        | NN2Vec + BLSTM-MIL     | DAIC-WOZ   | F1-score              | 0.8544              |
|                   |                                                                          |                         |            | Accuracy              | 96.7%               |
| Yang et al. [45]  | spectral LLD + cepstral LLD + prosodic LLD + voice quality LLD + functionals | DCNN-DNN                | DAIC-WOZ   | Female RMSE           | 5.669               |
|                   |                                                                          |                         |            | Female MAE            | 4.597               |
|                   |                                                                          |                         |            | Male RMSE             | 5.590               |
|                   |                                                                          |                         |            | Male MAE              | 5.107               |
| Jain [46]         | MFCC                                                                      | Capsule Network         | VCTK corpus | Accuracy              | 0.925               |
| Chao et al. [47]  | spectral LLD + cepstral LLD + prosodic LLD                               | LSTM-RNN                | AVEC2014   | unavailable           | unavailable         |
|                   |                                                                          |                         |            | unavailable           | unavailable         |
| Gupta et al. [48] | spectral LLD + cepstral LLD + prosodic LLD + voice quality LLD + functionals | DNN                     | AVID-Corpus | unavailable           | unavailable         |
| Kang et al. [50]  | spectral LLD + prosodic LLD + articulatory features                     | DNN                     | AVEC2014   | RMSE                  | 7.37                |
|                   |                                                                          |                         |            | MAE                   | 5.87                |
|                   |                                                                          |                         |            | Pearson’s Product Moment Correlation coefficient | 0.800               |
| Tzirakis et al. [58] | raw signal                                                              | CNN and 2-layers LSTM  | RECOLA     | loss function based on CCC | .440(arousal)       |
|                   |                                                                          |                         |            | .787(valence)         | .440(arousal)       |
| Tzirakis et al. [28] | raw signal                                                              | CNN and LSTM            | RECOLA     | CCC                   | .686(arousal)       |
|                   |                                                                          |                         |            |                       | .261(valence)       |
| Tzirakis et al. [41] | raw signal                                                              | CNN                     | RECOLA     | CCC                   | .699(arousal)       |
|                   |                                                                          |                         |            |                       | .311(valence)       |
| He et al. [21]    | LLDs + eGeMAPS                                                           | DBLSTM-RNN              | RECOLA     | CC                    | .836(arousal)       |
|                   |                                                                          |                         |            | RMSE                  | .529(valence)       |
|                   |                                                                          |                         |            | CCC                   | .099(arousal)       |
|                   |                                                                          |                         |            |                       | .104(valence)       |
|                   |                                                                          |                         |            |                       | .800(arousal)       |
|                   |                                                                          |                         |            |                       | .398(valence)       |
Fig. 1: The diagram of the proposed approach. Three deep architectures are proposed

signal is computed as a sequence of Fast Fourier Transform (FFT) of windowed audio segments. The audio signal is split into 256 segments and the spectrum of each segment is computed. The Hanning window is applied to each segment. The spectrogram plot is a color image of $1900 \times 1200 \times 3$. The image is resized to $224 \times 224 \times 3$ before being fed to the High-level features descriptor.

High-Level features descriptor. It is a deep CNN, it takes as input the spectrogram of the audio signal. Its architecture, as shown in Fig. 1, is composed by two same blocks of layers. Each block is composed of a two-dimensional (2D) convolutional layer followed by a ReLU activation function, a second convolutional layer, a ReLU, a dropout and max pooling layer, a third convolutional layer and last ReLU activation function.

3) MFCC-based CNN architecture: Like the spectrogram-based CNN, the MFCC-based CNN presents a low-level followed by high-level features descriptors. Low-level audio features descriptor is applied and the extracted features are fed to convolutional neural network to extract high-level features (see Fig. 1).

Low-level features descriptor MFCC. The speech signal is first divided into frames by applying a windowing function of 2.5s at fixed intervals of 500 ms. The Hamming window is used as window function to remove edge effects. A cepstral feature vector is then generated for each frame. The Discrete Fourier Transform (DFT) is computed for each frame. Only the logarithm of the amplitude spectrum is retained. The spectrum is after smoothed to emphasize perceptually meaningful frequencies. 24 spectral components into 44100 frequency bins are collected in the Mel frequency scale. The components of the Mel-spectral vectors calculated for each frame are highly correlated. Therefore, the Karhunen-Loeve (KL) transform is applied to the Mel-spectral vectors to decorrelate their components. The KL transform is approximated by the Discrete Cosine Transform (DCT). Finally, 177 cepstral features are obtained for each frame.

High-Level features descriptor. The MFCC features are fed to a convolutional neural network to extract more complex and high-level features. Many parameters may affect the performance of the CNN such as the number of layers. To avoid overfitting problem, only two one-dimensional (1D) convolutional layers followed by a ReLU activation function each are performed.

4) EmoAudioNet: Fusion of spectrogram-based and MFCC-based CNNs: The third architecture called EmoAudioNet is the fusion of the spectrogram-based CNN and the MFCC-based CNN as shown in Fig. 1. Combining the responses of the two deep CNNs allows to study simultaneously the time-frequency representation and the texture-like time frequency representation of the audio signal. The output of the spectrogram-based CNN is a feature vector of size 1152, while the output of the MFCC-based CNN is a feature vector of size 2816. The responses of the two networks are fused in a fully connected layer in order to generate the label prediction for nine emotion levels.

5) Data augmentation: a data augmentation step is considered for many reasons: to overcome the problem of data scarcity by increasing the quantity of training data and then
avoid overfitting problem, to explore different audio deforma-
tions and their impact on the model’s performance and also to
improve the robustness of the model to noise.

It has been demonstrated that adding noise to audio signals
in the training improves the robustness of speech recognition
approaches. Two different types of audio augmentation tech-
niques are considered in this work by perturbing the audio
signal:

- **Adding noise**: mix the audio signal with random noise.
  Each mix \( z \) is generated using \( z = x + \alpha \times \text{rand}(x) \)
  where \( x \) is the audio signal and \( \alpha \) is the noise factor. In
  our experiments, \( \alpha = 0.01, 0.02 \) and \( 0.03 \).
- **Pitch Shifting**: lower the pitch of the audio sample (while
  keeping the duration unchanged). Each sample was pitch
  shifted by 3 values (in semitones): 0.5, 2, 5

Each perturbation is applied directly to the audio signal prior
to using it as input data to train the network.

6) **Transferring Knowledge from audio related task**: Deep
Neural networks tend to learn general features, from the first
layer, which are not specific to a particular dataset or task. Meanwhile, the features computed in the last layer are more
specific and depend greatly on the chosen dataset and task.
Initializing a network with transferred features from almost
any number of layers can produce a boost to generalization
even after fine-tuning to the target dataset \([55]\).

Thus, in this work, knowledge is transferred from audio-
related task by first pre-training EmoAudioNet to solve another
task and second, fine-tuning the pre-trained network on speech
emotion classification dataset. This first initialization leverage
and transfer the knowledge from related task domain to
emotion classification domain. An evaluation framework of
knowledge transfer is performed and two-stepped fine-tuning
strategy is developed to transfer knowledge from an auxiliary
dataset:

- pre-training step: the EmoAudioNet is first trained by a
  related task that owns enough labeled audio data,
- fine-tuning step: the parameters learnt in the pre-training
  step are used as initialization for new task.

IV. EXPERIMENTS AND RESULTS

A. Datasets

Several publicly available datasets are used in our experi-
ments to evaluate the performances of the proposed architecture
on affect recognition and on depression identification and
assessment:

- **Dataset for affect recognition experiments**: RECOLA
dataset \([52]\) is used in our experiments to assess emotion
  and affect from speech. It is a multimodal corpus of
affective interactions in French. 46 subjects participate to
data recordings which include multimodal data (audio,
video, ECG and EDA) recorded continuously and
synchronously. Only 23 audio recordings of 5 minutes
of interaction are made publicly available and used in
our experiments. Participants engaged in a remote
discussion according to a survival task. In addition
to their recordings, six annotators measured emotion
continuously on two dimensions: valence and arousal.

- **Dataset for depression recognition and assessment
experiments**: DAIC-WOZ depression dataset \([54]\) is
  used in our experiments to assess depression from
  speech recordings. It is introduced in the AVEC2017
  challenge \([11]\) and it is a part of the large corpus
DAIC. The Distress Analysis Interview Corpus (DAIC)
contains clinical interviews aimed to study the diagnosis
of psychological distress conditions: depression, post
traumatic stress disorder (PTSD), etc. It provides
audio recordings of 189 clinical interviews of 189
participants answering the questions of an animated
virtual interviewer named Ellie. Each recording is
labeled by the PHQ-8 score and the PHQ-8 binary. The
PHQ-8 score defines the severity level of depression of
the participant and the PHQ-8 binary defines whether
the participant is depressed or not. For technical reasons,
only 182 audio recordings are used. The average length
of the recordings is 15 minutes with a fixed sampling
rate of 16 kHz.

- **Dataset for transfer learning experiments**: LibriSpeech
dataset \([53]\) is the dataset used to pre-train the EmoAu-
dioNet network before fine-tuning it with RECOLA
dataset. LibriSpeech is public corpus of read English
speech based on public domain audio books and It
contains 1000 hours of speech sampled at 16 kHz. In
our experiments, a subset of this corpus is used to pre-
train the proposed model. In this subset, the speech of 40
speakers among them 20 males and 20 female is recorded.
For each speaker, approximately, eight minutes of speech
are recorded, for a total of approximately 5 hours and 20
minutes.

B. Experimental Setup

**Spectrogram-based CNN architecture**: The number of
channels of the convolutional and pooling layers are both 128.
While their filter size is \( 3 \times 3 \). RELU is used as activation
function for all the layers. The stride of the max pooling is 8.
The dropout fraction is 0.1. The output layer is dense layer
of size 10 with a Softmax activation function.

**MFCC-based CNN architecture**: The input of this
network is one-dimensional and of size \( 177 \times 1 \). The filter
size of its two convolutional layers is \( 5 \times 1 \). RELU is used
as activation function for all the layers. The dropout fraction
is 0.1 and the stride of the max pooling is 8. Likewise the
spectrogram-based CNN, the output layer is dense layer
of size 10 with a Softmax activation function.

**EmoAudioNet architecture**: fuses the responses of the
spectrogram-based and the MFCC-based CNNs. The two
features vectors are concatenated and fed to a fully connected
layer of 10 neurons. A Softmax function is used as an
activation function of the fully connected layer.
RMSProp optimizer with its default configuration is used in the training of the three architectures. The training rate is set experimentally to 1e-5. This rate is reduced when the loss value stops decreasing. The batch size is fixed to 100 samples. The number of epochs for training is set to 500. An early stopping is performed when the accuracy stops improving after 10 epochs. The train-test split was performed using a stratified approach that ensures the same distribution of classes in each of the training and testing sets. RECOLA dataset is randomly divided into 80% for training and 20% for testing. During the training phase, 90% of the training set is used for learning the weights and 10% is used for validation.

C. Experimental results on spontaneous and continuous emotion recognition from speech

1) Results of three proposed CNN architectures: Time-Continuous prediction of natural and spontaneous emotions in two dimensions (arousal and valence) is investigated on RECOLA dataset. We analyze the experimental results of the three proposed architectures: the MFCC-based CNN, the spectrogram-based CNN and the EmoAudioNet network. Table II and Table IV display the results of prediction of arousal and valence respectively. We can see that EmoAudioNet outperforms MFCC-based CNN and the spectrogram-based CNN with an accuracy of 89% and 91% for predicting arousal and valence respectively. The accuracy of the MFCC-based CNN is around 70% and 71% for arousal and valence respectively. The spectrogram-based CNN is slightly better than the MFCC-based CNN and its accuracy is 75% for predicting arousal and 73% for predicting valence.

On the development set, the accuracy of EmoAudioNet is up to 94% for arousal and 95% for valence. On the test set, the accuracy exceeds 89% for arousal and 91% for valence. Thus, EmoAudioNet can accurately predict arousal and valence with an accuracy of approximately 90% when the data is acquired in close conditions. Others metrics are considered to evaluate the performances of EmoAudioNet: Pearson Coefficient Correlation (CC) and the Root Mean Square of Error (RMSE). EmoAudioNet has a CC of 0.90 for predicting arousal and 0.92 for predicting valence, and has also a small Root Mean Square of Error (RMSE) of 0.11 for arousal’s prediction and 0.15 for valence’s prediction.

2) Results of Transferring knowledge from other task: In our experiments, EmoAudioNet is pre-trained on LibriSpeech dataset to solve gender classification. This first initialization leverages and transfers knowledge from gender classification domain to Emotion recognition domain. As shown in Table III pre-training EmoAudioNet on Librispeech dataset and fine-tuning on RECOLA dataset, increases slightly the performances of predicting arousal and the accuracy reaches 90.37% on the test set, while the CC and the RMSE achieve 0.9156 and 0.1180, respectively. However, transferring knowledge from gender classification to valence level prediction is not efficient and even decreases the performance of EmoAudioNet.

Pre-training EmoAudioNet over an independent task has improved its performances in predicting arousal level from speech even without balancing the input dataset or increasing the number of depressed participants. EmoAudioNet has learned more abstract and complex features in the first layers when it is pretrained on gender classification which improves its performances on time-continuous prediction of natural and spontaneous emotions in the arousal dimension. For the valence dimension, it is important to investigate in future work the performance of EmoAudioNet when it is pre-trained on more complex and more large dataset and more related task.

3) Comparisons of EmoAudioNet and the state-of-the art methods for arousal and valence prediction on RECOLA dataset: In this section, we compare the performance of EmoAudioNet architecture with state-of-the-art methods on the RECOLA database. EmoAudioNet model pre-trained on LibriSpeech dataset and fine-tuned on RECOLA dataset has the best Pearson’s Correlation Coefficient (CC) of 0.9156 for arousal prediction. In term of the Root Mean Square of Error (RMSE), the approach proposed by He et al. [21] outperforms all the existing methods with an RMSE equal to 0.099. For valence prediction, EmoAudioNet outperforms other methods with a CC equal to 0.9221 while He et al. [21] has the best RMSE of 0.104.

D. Experimental results on automatic clinical depression recognition and assessment

EmoAudioNet framework is evaluated on two tasks on the DAIC-WOZ corpus. The first task is to predict depression from speech under the PHQ-8 binary test. The second task concerns the prediction of depression severity levels from speech under the PHQ-8 scores test.

1) EmoAudioNet performances on depression recognition task: EmoAudioNet is trained to predict the PHQ-8 binary test where a score of 0 identifies non-depression and a score of 1 identifies depression. Our proposed architecture performances are summarized in Table V and in Fig. 2a. The overall accuracy achieved in predicting depression reaches 73.25% with a root mean square error of 0.467. On the test set, 60.52% of the samples are correctly labeled with non-depression, whereas, only 12.73% are correctly diagnosed with depression. The low rate of correct classification of non-depression can be explained by the imbalance of the input data on the DAIC-WOZ dataset and the small amount of the participants labeled as depressed.

The accuracy is not the best metric to use when working with an imbalanced dataset. F1 score is one the metrics that have been designed to deal with the non-uniform distribution of class labels by giving a weighted average of precision and recall. The non-depression F1 score reaches 82% while the depression F1 score reaches 49%. Almost half of the samples
TABLE III: RECOLA dataset results for prediction of arousal. The results obtained for the development and the test sets in term of three metrics: the accuracy, the Pearson’s Coefficient Correlation (CC) and the Root Mean Square error (RMSE). (*) EmoAudioNet is pre-trained on LibriSpeech dataset and fine-tuned on RECOLA dataset.

| Experience            | Development | Test |
|-----------------------|-------------|------|
|                       | Accuracy | CC | RMSE | Accuracy | CC | RMSE |
| MFCC-based CNN         | 81.93%   | 0.8130 | 0.1501 | 70.23% | 0.6981 | 0.2065 |
| Spectrogram-based CNN  | 80.20%   | 0.8157 | 0.1314 | 75.65% | 0.7673 | 0.2099 |
| EmoAudioNet            | 94.49%   | 0.9521 | 0.0082 | 89.30% | 0.9069 | 0.1229 |
| EmoAudioNet (*)        | 95.16%   | 0.9555 | 0.07895 | 90.37% | 0.9156 | 0.1180 |

TABLE IV: RECOLA dataset results for prediction of valence. The results obtained for the development and the test sets in term of three metrics: the accuracy, the Pearson’s Coefficient Correlation (CC) and the Root Mean Square error (RMSE). (*) EmoAudioNet is pre-trained on LibriSpeech dataset and fine-tuned on RECOLA dataset.

| Experience            | Development | Test |
|-----------------------|-------------|------|
|                       | Accuracy | CC | RMSE | Accuracy | CC | RMSE |
| MFCC-based CNN         | 83.37%   | 0.8289 | 0.1405 | 71.12% | 0.6965 | 0.2082 |
| Spectrogram-based CNN  | 78.32%   | 0.7984 | 0.1446 | 73.81% | 0.7598 | 0.2132 |
| EmoAudioNet            | 95.42%   | 0.9568 | 0.0625 | 91.44% | 0.9221 | 0.1118 |
| EmoAudioNet (*)        | 87.55%   | 0.9028 | 0.1222 | 83.07% | 0.8624 | 0.1508 |

predicted with depression are correctly classified with a precision of 51.71%. The number of non-depression samples is twice the number of samples labeled with depression. Thus, adding more samples of depressed participants would significantly increase the model’s ability to recognize depression and to correctly identify samples with depression.

2) EmoAudioNet performances on depression severity levels prediction task: The depression severity levels are assessed by the PHQ-8 scores ranging from 0 to 23. Non-depression is identified with a score of 0, whereas a severe depression is diagnosed when the PHQ-8 score reaches maximum scores. The EmoAudioNet is trained to predict the 24 classes of the depression severity levels. The root mean square error achieved when predicting the PHQ-8 scores is 2.6 times better than the RMSE achieved with the depression recognition task. The test loss reaches 0.18 compared to a 0.1 root mean square of error on the training set.

3) Results of Transferring knowledge from another task: The number of samples labeled with depression is relatively small. To overcome this problem, an experiment of transferring knowledge from other task is performed. Thus, the EmoAudioNet is pre-trained on LibriSpeech, a dataset for gender classification. The weights of the pre-trained model are saved and re-used as a starting point for the depression recognition task using the DAIC-WOZ corpus.

In this experiment, the model’s overall accuracy improved by 0.88% to reach 74.13% on the test set. The Table. VI summarizes the EmoAudioNet performances where the test CC increased by 0.02 to reach 0.5 and the test loss or RMSE reached 0.47. Both, the non-depression and the depression F1 scores improved by 1% to reach 83% and 50%, respectively. This improvement is explained by the increase in the recall and precision rates of both classes as shown in the confusion matrix in Fig. 2b. The recall of the class non-depression increased by 2.27% while the precision of depression increased by 4.57%. After performing transfer learning, nearly, 13% of the samples were correctly labeled with depression on the test set. More than the half of these samples were correctly classified with a precision of 56.28%.

These results show that EmoAudioNet has gained more knowledge in diagnosing samples with depression without having to add more input data. Pre-training it on the LibriSpeech dataset, has ameliorated its ability to learn more complex features leading it to perform better in the depression recognition task.
TABLE V: EmoAudioNet Performances on the Development and Test sets in the Depression Assessment Task

| Results     | Accuracy | CC  | RMSE |
|-------------|----------|-----|------|
| Development Set | 81.96%   | 0.601 | 0.394 |
| Test Set     | 73.25%   | 0.482 | 0.467 |

TABLE VI: EmoAudioNet Performances on the Development and Test Sets in the Depression Assessment Task after performing Transfer of Knowledge

| Results     | Accuracy | CC  | RMSE |
|-------------|----------|-----|------|
| Development Set | 86.21%   | 0.642 | 0.60  |
| Test Set     | 74.13%   | 0.507 | 0.47  |

4) Comparisons of EmoAudioNet and the state-of-the-art methods for depression prediction on DAIC-WOZ dataset:

Table VIII compares the performances of EmoAudioNet architecture with the state-of-the-art approaches and shows the best performing approaches evaluated on the DAIC-WOZ dataset. To the best of our knowledge, the best performing approach in the literature on DAIC-WOZ dataset is the proposed approach in [44] with an F1 score of 85.44% and an accuracy of 96.7% . The proposed NN2Vec features with BLSTM-MIL classifier achieves this good performance thanks to the leave-one-speaker out cross-validation approach. Comparing to the other proposed approaches where a simple train-test split is performed, giving the model the opportunity to train on multiple train-test splits increase the model performances especially in small datasets.

In the depression recognition task, the EmoAudioNet outperforms the proposed architecture in [56] based on a Convolutional Neural Network followed by a Long Short-Term Memory network. The non-depression F1 score achieved with EmoAudioNet is better than the latter by 13% with the exact same depression F1 score (50%).

Moreover, the EmoAudioNet outperforms the Long Short-Term Memory network in [57] in correctly classifying samples of depression. The depression F1 score achieved with EmoAudioNet is higher than the MFCC-based RNN by 4%. Meanwhile, the overall accuracy and loss achieved by the proposed MFCC-based RNN are better than EmoAudioNet by 2.14% and 0.07 respectively.

According to the summarized results of previous works in Table VIII the best results achieved so far in the depression severity level prediction task are obtained in [57]. The best root mean square error is achieved with the Long Short-Term Memory network to reach 0.168. The EmoAudioNet reaches almost the same loss with a very low difference of 0.012. Our proposed architecture outperforms the rest of the results in the literature in predicting the PHQ-8 scores. Both works in [16] and [45] propose an architecture based on a Deep Convolutional Neural Network combined with a Deep Neural Network (DCNN-DNN). Among these two studies, the best root mean square error is achieved in [16] over a group of depressed men. The loss achieved with EmoAudioNet is eight times better than the latter.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new emotion and affect recognition methods from speech approach based on deep neural networks called EmoAudioNet. It is the aggregation of an MFCC-based CNN and a spectrogram-based CNN, which study the time-frequency representation and the visual representation of the spectrum of frequencies of the audio signal. EmoAudioNet outperforms state-of-art approaches of automatic depression recognition and continuous dimensional affect recognition from speech on RECOLA and DAIC-WOZ databases. We have shown that transferring knowledge from other task improves the performance of EmoAudioNet on the target task. In future work, we are planning to test EmoAudioNet on others speech recognition applications and generalize it to other speech-related applications.

REFERENCES

[1] GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015," Lancet, 388, vol. 388, no 10053, pp. 1545-1602, 2015.
[2] Les chiffres et statistiques de la dépression en France et dans le monde. Retrieved 2019, Mai from https://www.la-depression.org/comprendre-la-depression/la-depression-en-chiffre/.
[3] The National Institute of Mental Health. (2018). Depression. Retrieved 2019, June 17 from https://www.nimh.nih.gov/health/topics/depression/index.shtml.
[4] info-depression.fr. Retrieved 2019, April from http://www.info-depression.fr/.
[5] P. F. Sullivan, M. C. Neale, and K. S. Kendler, "Genetic epidemiology of major depression: review and meta-analysis," The American Journal of Psychiatry, vol. 157, no. 10, pp. 1552-62, 2000.
[6] K. Ma, "Attachment theory in adult psychiatry. Part 1: Conceptualisations, measurement and clinical research findings," Advances in Psychiatric Treatment, vol. 12, pp. 440-449, 2006.
[7] G. M. Slavich, "Deconstructing depression: A diathesis-stress perspective (Opinion)," APS Observer, vol. 17, no. 9, 2004.
### TABLE VII: Comparisons of EmoAudioNet and the state-of-the-art methods for arousal and valence prediction on RECOLA dataset. The results obtained in term of three metrics: the accuracy, the Pearson’s Coefficient Correlation (CC) and the Root Mean Square error (RMSE). (*) EmoAudioNet is pre-trained on LibriSpeech dataset and fine-tuned on RECOLA dataset.

| Method                  | Arousal CC | Arousal RMSE | Valence CC | Valence RMSE |
|-------------------------|------------|--------------|------------|--------------|
| He et al. [21]          | 0.836      | 0.099        | 0.529      | 0.104        |
| Ringeval et al. [20]    | 0.322      | 0.173        | 0.144      | 0.127        |
| EmoAudioNet             | 0.9069     | 0.1229       | 0.9221     | 0.1118       |
| EmoAudioNet (*)         | 0.9156     | 0.1180       | 0.8624     | 0.1508       |

### TABLE VIII: Comparisons of EmoAudioNet and the state-of-the-art methods for prediction of depression on DAIC-WOZ dataset. The results obtained in term of three metrics: the accuracy and the Root Mean Square error (RMSE). (*) EmoAudioNet is pre-trained on LibriSpeech dataset and fine-tuned on DAIC-WOZ dataset. (**) The results of the depression severity level prediction task.

| Method                  | Accuracy | RMSE | F1 Score |
|-------------------------|----------|------|----------|
| Yang et al. [16]        | -        | 1.46 (**) (depressed male) | - |
| Yang et al. [45]        | -        | 5.59 (**) (male) | - |
| Valstar et al. [10]     | -        | 7.78 (**) | - |
| Al Hanai et al. [42]    | -        | 10.03 | - |
| Salekin et al. [44]     | 96.7%    | -    | 85.44%   |
| Ma et al. [56]          | -        | -    | 70% for non-depression and 50% for depression |
| Rejaibi et al. [57]     | 76.27%   | 0.4  | 85% for non-depression and 46% for depression |
| -                       | -        | 0.168 (**) | - |
| EmoAudioNet             | 73.25%   | 0.467| 82% for non-depression and 49% for depression |
| -                       | -        | 0.18 (**) | - |
| EmoAudioNet (*)         | 74.13%   | 0.47 | 83% for non-depression and 50% for depression |

---

[8] B.F. Jeronimus, R. Kotov, H. Riese, and J. Ormel, "Neuroticism’s prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: a meta-analysis on 59 longitudinal/prospective studies with 443 313 participants," Psychological Medicine, vol. 46, no. 14, pp. 2883-2906, 2016.

[9] A. Jan, H. Meng, Y.F.B.A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 3, pp. 668-680, 2017.

[10] 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 the 6th international workshop on audio/visual emotion challenge, ACM, pp. 3-10, 2016.

[11] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, pp. 3-9, 2017.

[12] C.M. Vazakopoulou, A. Pampouchidou, F. Yang, F. Meriaudeau, K. Marias, and M. Tsiknakis, "Dtection de la dpression par lanalyse de la gomtrie faciale et apprentissage automatique," Gretsi, 2017.

[13] H. Jiang, B. Hu, Z. Liu, G. Wang, L. Zhang, X. Li, and H. Kang, "Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features," Computational and mathematical methods in medicine., vol. 2018, 2018.

[14] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, T. Gedeon, M. Breakspear, and G. Parker, “A comparative study of different classifiers for detecting depression from spontaneous speech,” In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8022-8026, 2013.

[15] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhiya, S. Schniede, R. Cowie, and M. Pantic, “AVEC 2013: the continuous audiovisual emotion and depression recognition challenge,” In Proceedings of the 3rd ACM international workshop on Audio/Visual emotion challenge, ACM, pp. 3-10, 2013.

[16] L. Yang, H. Sahli, X. Xia, E. Pei, M.C. Oveneke and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, pp. 45-51, 2017.

[17] Z. Liu, B. Hu, L. Yan, T. Wang, F. Liu, X. Li, and H. Kang, "Detection of depression in speech," In 2015 international conference on affective computing and intelligent interaction (ACII), IEEE, pp. 743-747, 2015.
### Table: Confusion Matrices on DAIC-WOZ Test Set

|          | Non-Depression | Depression |          | Non-Depression | Depression |          |
|----------|----------------|------------|----------|----------------|------------|----------|
| **Actual** |                |            | **Precision** |                |            |          |
| **Non-Depression** | 1441 | 354        | 1795 | 80.28% | 19.72% |          |
| **Depression**    | 283       | 303        | 586     | 51.71% | 48.29% |          |
| **Recall**        | 1724      | 657        | 2381    | 83.58% | 46.12% | 73.25% |
|                   | 85.12%    | 53.88%     | 26.75%  |          |          |          |

(a) Training and testing EmoAudioNet on DAIC-WOZ

|          | Non-Depression | Depression |          | Non-Depression | Depression |          |
|----------|----------------|------------|----------|----------------|------------|----------|
| **Actual** |                |            | **Precision** |                |            |          |
| **Non-Depression** | 1456 | 376        | 1832 | 79.48% | 20.52% |          |
| **Depression**    | 240       | 309        | 549     | 56.28% | 43.72% |          |
| **Recall**        | 1696      | 685        | 2381    | 85.85% | 45.11% | 74.13% |
|                   | 74.11%    | 54.89%     | 25.87%  |          |          |          |

(b) Pre-training on LibriSpeech and after training and testing on DAIC-WOZ

---

[18] N. Cummins, J. Epps, M. Breakspear and R. Goecke, "An investigation of depressed speech detection: Features and normalization," In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[19] P. Lopez-Otero, L. Dacia-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for depression detection," In *Second International Workshop on Biometrics and Forensics*, IEEE, pp. 1-6, 2014.

[20] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ACM, pp. 3-8, 2015.

[21] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ACM, pp. 73-80, 2015.

[22] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. ifti, H. Gle, A.A. Salah, and M. Pantic, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ACM, pp. 3-13, 2018.

[23] A. Dhall, O.V. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, ACM, pp. 423-426, 2015.

[24] S. Haq, P.J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition," In *AVSP*. pp. 53-58, 2009.

[25] L.S.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, and N.B. Allen, "Detection of clinical depression in adolescents’ speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 574-586, 2010.

[26] M. Valstar, B.W. Schuller, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: The 4th international audio/visual emotion challenge and workshop," In *Proceedings of the 22nd ACM international conference on Multimedia*, ACM, pp. 1243-1244, 2014.

[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 *Proceedings of the 3rd ACM international workshop on Audio/Visual emotion challenge*, ACM, pp. 21-30, 2013.

[28] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adues features? end-to-end speech emotion recognition using a deep convolutional recurrent recurrent network," In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp. 5200-5204, 2016.

[29] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22-30, 2015.

[30] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," In *Proceedings of the 15th ACM on International conference on multimodal interaction*, ACM, pp. 509-516, 2013.

[31] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015: The 5th international audiovisual emotion challenge and workshop," In *Proceedings of the 23rd ACM international conference on Multimedia*, ACM, pp. 1335-1336, 2015.

[32] C.S. Ang, A. Bobrowicz, D.J. Schiano and B. Nardi, "Data in the wild: Some reflections," *Interactions*, vol. 20, no. 2, pp. 39-43, 2013.

[33] Z. Lian, Y. Li, J. Tao, and J. Huang, "Investigation of Multimodal Features, Classifiers and Fusion Methods for Emotion Recognition," arXiv preprint arXiv:1809.06225.

[34] J.H. Jeans, "Science & Music", *Courier Corporation*, 1968.

[35] N.Y. Melville, Standards Secretariat, *Acoustical Society of America*. ANSI S3, pp. 2-1989.

[36] G. Lindsey, "The Vowel Space," 2013.

[37] E. Kvedalen, "Signal processing using the Teager energy operator and other nonlinear operators," *Master, University of Oslo Department of Informatics*, vol. 21, 2003.

[38] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. Andr, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, and K.P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190-202, 2016.

[39] J. Kossaifi, R. Valecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," arXiv preprint arXiv:1901.02839.

[40] D. Kollias, P. Tzirakis, M.A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, pp. 1-23, 2019.

[41] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301-1309, 2017.

[42] T. Al Hanai, M.M. Ghassemi, and J.R. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," In *Interspeech*, pp. 1716-1720, 2018.

[43] S. Dham, A. Sharma and A. Dhall, "Depression scale recognition from audio, visual and text analysis," arXiv preprint arXiv:1709.05865.

[44] A. Salekin, J.W. Eberle, J. Glenn, B.A. Teachman, and J.F. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 2, p. 81, 2018.
[45] L. Yang, D. Jiang, X. Xia, E. Pei, M.C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, pp. 53-59, 2017.

[46] R. Jain, "Improving performance and inference on audio classification tasks using capsule networks," arXiv preprint arXiv:1902.05069, 2019.

[47] L. Chao, J. Tao, M. Yang, and Y. Li, “Multi task sequence learning for depression scale prediction from video,” In 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, pp. 526-531, 2015.

[48] R. Gupta, S. Sahu, C.Y. Espy-Wilson, S.S. Narayanan, "An Affect Prediction Approach Through Depression Severity Parameter Incorporation in Neural Networks," In Interspeech, pp. 3122-3126, 2017.

[49] L. Yang, H. Sahli, X. Xia, E. Pei, M.C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, pp. 45-51, 2017.

[50] Y. Kang, X. Jiang, Y. Yin, Y. Shang, and X. Zhou, X. "Deep transformation learning for depression diagnosis from facial images," In Chinese Conference on Biometric Recognition, Springer, Cham, pp. 13-22, 2017.

[51] G. Yu, and J.J. Slotine, J. J. "Audio classification from time-frequency texture," In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp. 1677-1680, 2009.

[52] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, D. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, pp. 1-8, 2013.

[53] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 5206-5210, 2015.

[54] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D.R. Traum, S. Rizzo, and L.P. Morency, "The distress analysis interview corpus of human and computer interviews," LREC, pp. 3123-3128, 2014.

[55] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks? " In Advances in neural information processing systems, pp. 3320-3328, 2014.

[56] 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, pp. 35-42, 2016.

[57] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech." arXiv preprint arXiv:1909.07208 2019.

[58] P. Tzirakis, J. Zhang, and B.W. Schuller, "End-to-end speech emotion recognition using deep neural networks," In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089-5093, 2018.