Detecting Depression in Less Than 10 Seconds: Impact of Speaking Time on Depression Detection Sensitivity

Nujud Aloshban
University of Glasgow (UK)
Nujud.Aloshban@glasgow.ac.uk

Anna Esposito
Universitá della Campania "Luigi Vanvitelli" (Italy)
Anna.Esposito@unicampania.it

Alessandro Vinciarelli
University of Glasgow (UK)
Alessandro.Vinciarelli@glasgow.ac.uk

ABSTRACT

This article investigates whether it is possible to detect depression using less than 10 seconds of speech. The experiments have involved 59 participants (including 29 that have been diagnosed with depression by a professional psychiatrist) and are based on a multimodal approach that jointly models linguistic (what people say) and acoustic (how people say it) aspects of speech using four different strategies for the fusion of multiple data streams. On average, every interview has lasted for 242.2 seconds, but the results show that 10 seconds or less are sufficient to achieve the same level of recall (roughly 70%) observed after using the entire inteview of every participant. In other words, it is possible to maintain the same level of sensitivity (the name of recall in clinical settings) while reducing by 95%, on average, the amount of time requireed to collect the necessary data.

CCS CONCEPTS

• **Human-centered computing**; • **Applied computing** → **Health informatics**;

KEYWORDS

Depression Detection; Computational Paralinguistics; Social Signal Processing; Speech, Language

ACM Reference Format:
Nujud Aloshban, Anna Esposito, and Alessandro Vinciarelli. 2020. Detecting Depression in Less Than 10 Seconds: Impact of Speaking Time on Depression Detection Sensitivity. In *2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual Event, Netherlands*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

According to the World Health Organisation, "*at a global level, over 300 million people are estimated to suffer from depression, equivalent to 4.4% of the worlds population [...] the single largest contributor to global disability (7.5% of all years lived with disability in 2015) [...] the major contributor to suicide deaths, which number close to 800,000 per year.*" [41]. Still, diagnosing depression remains a challenging problem because the discrimination between the pathology and ordinary forms of stress, anxiety or sadness is difficult [40]. In such a context, Artificial Intelligence can help clinicians through the development of automatic approaches for the identification of people actually affected by depression.

To the best of our knowledge, the computing efforts made so far have targeted mainly the improvement of the detection performance and have addressed only to a limited extent, if at all, the problem of how much data is necessary to make a reliable decision about an individual (see Section 2). For this reason, this article investigates whether it is possible to perform depression detection with 10 seconds of speech or less and, if yes, to what extent. The main reason why such a problem is important is that realistic application scenarios require one to deal with recordings that contain only a few words (e.g., the use of data collected at help lines [21]). Furthermore, when the speech data is obtained through interviews or other forms of interaction that involve medical personnel, reducing the amount of time necessary to gather enough information lowers the costs associated to depression diagnosis.

The experiments are based on state-of-the-art methodologies for joint modeling of linguistic and acoustic aspects of speech (corresponding to *what* people say and *how* they say it, respectively). The tests have been peformed over a corpus of 59 clinical interviews that have involved 29 participants diagnosed with depression by professional psychiatrists and 30 that never experienced mental health issues. The results show that less than 10 seconds are sufficient to achieve the same level of recall (around 70%) that can be obtained when using the entire clinical interview of every participant. Since the average length of an interview is 242.2 seconds, such a result means that the time required to interview the participants can be reduced by 24 times without significant sensitivity losses (sensitivity is the way recall is referred to as in clinical settings).

The main reason why this is important is that recall measures the effectiveness at recognizing all depressed individuals as such, i.e., at avoiding type II errors (classifying a depressed individual as healthy), those that lead to the most negative consequences from a clinical point of view. In fact, in the case of a type I error (a control individual classified as depressed), the consequence is that a healthy individual will be examined

ICMI '20, October 25–29, 2020, Virtual Event, Netherlands

ACM ISBN 978-1-4503-7092-9/20/08. . . $15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

more thoroughly by doctors, but such an extra medical attention will not be of harm. In contrast, in the case of a type II error, a depressed individual will go undetected and will not undergo proper treatment, thus joining the estimated 79% of depression patients that do not receive appropriate care [24], a major issue in nowadays psychiatry.

The rest of this article is organised as follows: Section 2 surveys previous work, Section 3 describes the data, Section 4 presents the depression detection approach, Section 5 reports on experiments and results, and the final Section 6 draws some conclusions.

2 SURVEY OF PREVIOUS WORK

Coherently with its major impact on society (see Section 1), depression is the psychiatric problem that the computing community addresses more frequently. In particular, the pathology has been the target of at least four benchmarking campaigns based on two corpora used in a large fraction of works published in the last decade. The first corpus includes 292 people asked to perform a Human-Computer Interaction task [38, 39], the second involves more than 200 individuals interacting with an artificial agent [34, 37]. In both cases, the goal is the inference of the scores resulting fom the administration of self-assessment questionnaires, namely the BDI-II [8] and different versions of the *Patient Health Questionnaire* (PHQ) [18].

The works using the data above include approaches based on facial behaviour [2, 45, 46], paralinguistics [14] or multimodal combinations of different cues [43]. The experiments presented in [2] focus on temporal dynamics of facial expressions and predict the self-assessment scores resulting from the questionnaires mentioned above. The best result is a 9.2 Root Mean Square Error (RMSE). In [45], the goal is to identify facial depression markers, i.e., face regions and actions most likely to account for depression, while in [46], it is the inference of the BDI-II scores, a task performed with 9.8 RMSE. The approach proposed in [14] focuses on speech and includes two main steps, the first is the inference of the particular range of the BDI-II score a person falls in, and the second is the inference of the exact score in such a range. An RMSE of 8.2 is the best result that the article reports. The experiments in [43] are based on a multimodal approach modeling face behaviour, speech and the manual transcription of what people say. The best result is an F1 score of 75% in identifying people above the PHQ-8 threshold score corresponding to depression.

The experiments presented in [31] show that text analysis techniques allow one to identify social media posts written by people that claim to be depressed. However, it is not possible to test whether the claim is true. Actual depression detection, meaning that the people involved in the experiments have been diagnosed by a doctor, is the task addressed in [3, 4, 10, 19, 23, 44]. The analysis proposed in [44] shows that there is a statistically significant correlation between measurable aspects of speech and depression, while the other works propose approaches for detecting depression in different signals. The approach in [10] performs such a task with accuracy around 90% using Electro-Encephalograms (EEG). Such a performance is similar to the one reported in [3], where the accuracy is 88%, using a multimodal combination of paralinguistics, head pose and gaze. Such a work follows up on previous work based on head movements. The work reports an accuracy higher than 70% achieved over a subset of the Black Dog Corpus (30 control and 30 depressed) [4]. The approach proposed in [19] focuses on facial expressions and shows that changes in the way these are displayed correspond to the severity of depression. A last cue that has been taken into account is body movement (including gestures) that leads to an F1 measure of up to 80% in combination with head pose and facial expressions [23].

Like this article, several works have addressed depression detection or inference of self-assessment scores using speech and, possibly, its transcription. In the experiments presented in [21], the focus is on the use of mobile phones (characterised by low quality and noisy audio) and the need to use short utterances. The results show that an accuracy up to 72% can be achieved in identifying people with PHQ-9 scores higher than 9 (the depression threshold). The experiments in [13, 28] focus on the paralinguistic differences between depressed and non-depressed speakers. In [28], the experiments are performed over adolescents because their voice is not fully formed and, hence, it might not be as informative as in the case of adults. The results show that the feature allowing one to better discriminate between depressed speakers and the others is the energy (related to how loud someone speaks), especially when measured with the Teager Operator [36]. In the second work [13], the data show that non-depressed individuals tend to display higher variability in their way of speaking. The experiments presented in [33] show that it is possible to detect depressed speakers using MFCC and Recurrent Neural Networks.

Finally, several works have addressed the problem of combining speech and its transcritpion like this work. The experiments in [30] suggest that acoustic aspects, while being a valuable source of information, should not be used without taking into account transcritpions. In particular, the experiments show that it is the joint modeling of acoustic and linguistic aspects that leads to the best results. However, the results presented in [42] appear to suggest that the best F1 measure, even if by just one point (71% against 70%), results from the use of the sole transcriptions. Similarly, other experiments show that the joint modeling of paralinguistics and lexical choice leads to lower F1 measures than the use of lexical choice alone (69% against 67%). Still, the same work shows that the use of gating mechanisms can improve the performance of the combination (F1 measure 80%) [35]. The experiments in [1], based on joint modeling of speech and manual transcriptions, shows that the performance can be improved by taking into account when a sentence has been uttered during an interview. In this case, a multimodal approach based on speech and text leads to the best performance (F1 measure 77%).

Overall, the state-of-the-art suggests that none of the behavioural cues considered so far (facial expressions, language, gestures, etc.) clearly outperform the others. Furthermore, the application of multimodal approaches does not necessarily work, especially when considering linguistic and acoustic aspects of speech (see above). One possible explanation is that depression interplays with so many different factors (physiology, socio-economic status, age, gender, etc. [22]) that its detectable traces change considerably from one person to the other. As a solution, at least partial, this work includes the collecion of data that are as balanced as possible in terms of age, gender and education level (see Section 3). In this way, it is possible to limit the effect of factors other than depression.

3 DATA COLLECTION

The data used in the experiments of this work have been collected in three Mental Health Centres in Southern Italy. Table 1 provides information about gender, age and education level of the 59 participants involved in the experiments. Furthermore, the table shows that the participants can be split into two groups, namely *depression* (29 persons diagnosed with depression by professional psychiatrists) and *control* (30 persons that have never experienced any mental health issues).

The gender distribution is the same in both groups and, overall, the number of female participants is 2.47 times higher than the number of male ones. This reflects the tendency of women to develop depression roughly twice as frequently as men do [5]. In terms of age distribution, there is no difference between the two groups ($p << 0.01$ according to a two-tailed t-test) and the age range excludes children, adolescents and people above 70 because these tend to manifest depression less frequently [16, 25, 29]. For what concerns the education level, a two-tailed χ^2 test shows that the distribution is the same for both groups ($p < 0.05$). The balance in terms of gender, age and education level limits the possibility that observable differences between the two groups result from factors other than depression.

Every participant has been invited to participate in an interview in which an experimenter has posed always the same questions and always in the same order. The questions address aspects of everyday life (e.g., activities during the last week end) and the experimenters have limited their interventions to the minimum. This aims at ensuring the collection of the largest possible amount of spontaneous speech. As a result, the participants speak, on average, 90% of the interview time. However, there is a statistically significant difference between depression and control participants that speak, on average, 95.0% and 85.3% of the time, respectively ($p << 0.01$ according to a two-tailed t-test). The main probable reason is that control participants tend to involve the interviewer in a conversation, while depression ones answer the questions and do not try to interact further. The discrimination between depressed and non-depressed participants has been made by professional psychiatrists.

Group	F	M	Avg. Age	Age Range	L	H
Depression	21	8	45.7	23-69	16	13
Control	21	9	44.0	23-68	12	18
Total	42	17	44.8	23-69	28	31

Table 1: The table provides demographic information about the experiment participants. Acronyms F and M stand for female and male, respectively, while acronyms L and H stand for lower (up to 8 years of study) and higher (at least 13 years of study) education level, respectively.

On average, every interview lasts for 242.2 seconds, but there is a statistically significant difference between the average durations for the two groups, corresponding to 216.5 and 267.1 seconds for depression and control participants, respectively ($p < 0.01$ according to a one-tailed t-test). This is not surprising because the literature provides evidence that depressed individuals tend to engage less in social interactions and, therefore, to speak less than people that are not affected by the pathology [9, 17]. Every interview has been segmented into *clauses*, atomic linguistic structures that include a noun, a verb and a complement. These are meaningful analysis units that make it possible to analyse how the performance of a depression detection approach changes when the amount of data about a person increases. Likely because of the duration differences mentioned earlier, control participants utter an average of 127.0 clauses, while depression ones utter 103.3. According to a one-tailed t-test, such a difference is statistically significant ($p < 0.01$).

4 DEPRESSION DETECTION

Figure 1 shows the main components of the unimodal and multimodal recognition approaches used in this work, namely *encoding*, *multimodal representation*, *classification* and *aggregation*. In both unimodal and multimodal cases, the input corresponds to the N clauses $\{c_1, c_2, ..., c_N\}$ that a given participant has uttered (the value of N changes from one participant to the other). Each clause c_i is classified individually resulting into N individual outcomes $\{\hat{l}_1, \hat{l}_2, ..., \hat{l}_N\}$, where $\hat{l}_j$ is one of the two possible classes, i.e., *depression* or *control*. The final classification outcome is obtained by aggregating the $\hat{l}_j$s though a majority vote. In other words, a participant is assigned to the class her or his clauses are most frequently assigned to. The rest of this section describes encoding, multimodal representation and classification in detail (the aggregation corresponds to the majority vote and no further detail is provided).

4.1 Encoding

The encoding component includes two main steps, namely *feature extraction* and *unimodal representation*. Since every clause includes both an audio signal and its transcription, two distinct feature sets are extracted, one from the audio and the other from the text. In both cases, the result is a

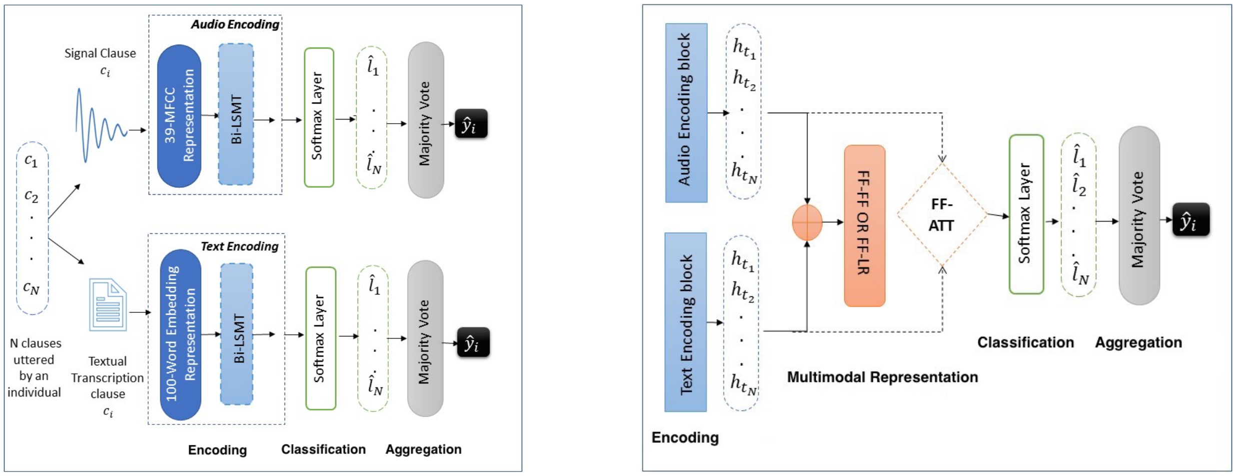

Figure 1: Left and right schemes show unimodal and multimodal recognition approach, respectively. In the unimodal case (left scheme), every clause is encoded through a process that includes two main steps, namely feature extraction and representation. The output of the encoding is fed to a softmax layer that provides a classification outcome for each clause. The individual outcomes are then aggregated through a majority vote. In the multimodal case, the output of the two encoding components (one per modality) are fed to a multimodal representation step. This takes three different forms: concatenation and Logistic Regression that performs a classification, concatenation and Feed Forward network, or Gated Multimodal Units. In the last two cases, the output of the multimodal representation is fed to a softmax layer that performs a classification. The final step is the aggregation of the individual clause classification outcomes.

sequence of feature vectors that are fed to a Bidirectional Long Short-Term Memory network (Bi-LSTM) [20] acting as an encoder.

4.1.1 Audio Feature Extraction. The feature extraction process segments the audio signal into 25 ms long analysis windows that start at regular time steps of 10 ms (two consecutive windows overlap by 15 ms). The length of the windows corresponds to the time-scale of acoustic phenomena in speech. Therefore, it is expected to ensure that the the signal properties remain relatively stable in the time intervals the windows correspond to. After the segmentation, the signal intervals enclosed in every window are mapped into feature vectors where the components are the first $F = 39$ *Mel Frequency Cepstral Coefficients* (MFCC) [36]. Correspondingly, each clause is converted into a sequence $A = (a_1, a_2, ..., a_{T_A})$, where a_k is the F-dimensional vector extracted from the k^{th} window. The value of T_A, the number of vectors allowed in A, is set through cross-validation during the experiments. As a result, every clause is mapped into a two-dimensional matrix $A \in \mathbb{R}^{T_a \times F}$.

4.1.2 Text Feature Extraction. The clause transcriptions are converted into sequences of vectors through *Word Embedding*, an approach that has been shown to capture linguistic and semantic characteristics of words, meaning that words similar along such dimensions tend to be mapped into similar vectors [11]. In particular, the transcription of every clause is mapped into a sequence $S = (s_1, s_2, ..., s_{T_S})$, where s_k is a D-dimensional vector corresponding to the k^{th} word and T_S is the maximum number of vectors allowed in S. The value of D has been set to 100 (no other values have been tried), while the value of T_S has been set through cross-validation during the experiments. As a result, S is represented as as a two-dimensional matrix $S \in \mathbb{R}^{T_S \times D}$.

The Word Embedding approach used in the experiments is static, i.e., it maps every word always into the same vector, irrespectively of the different contexts in which it appears. More recent approaches (e.g., the *Bi-Directional Encoder Representations from Transformers* [15]) allow one to overcome such a limitation, but they did not lead to any improvement in the experiments of this work. The probable reason is that clauses tend to be short (3.9 words on average) and, therefore, the contextual information is insufficient for this particular type of text. For this reason, the experiments of this work make use of a static Word Embedding approach.

4.1.3 Unimodal Representation. After the feature extraction process, the clauses are mapped into sequences of feature vectors $X = (x_1, x_2, ..., x_T)$, where X corresponds to A or S and T corresponds to T_A or T_S depending on wheter the features have been extracted from the audio signal or from its transcription (see Sections 4.1.1 and 4.1.2). The main motivation is that the input data is sequential and, in particular, audio vectors a_k correspond to different points in time of the speech signal, while vectors s_k correspond to different words in a text. However, the vectors do not carry sequential information, i.e., they do not encode possible relationships between feature vectors extracted at different points in time. For this reason, the X sequences are fed to Bi-LSTMs [20], well known to capture such relationships, if any.

4.2 Multimodal Representation

The multimodal combination approach builds upon the unimodal representations introduced in Section 4.1 and implements different strategies for the combination of lexical and paralinguistic information extracted from the data. The main reason for using a wide spectrum of approaches is to ensure that the results of the experiments do not depend on a particular approach being used, but correspond to the actual information in the data. The rest of this section presents every multimodal combination approach in detail.

4.2.1 Late Fusion (LF). The classification of the unimodal representations takes place by feeding the output of the encoders to a softmax layer trained to minimize the cross-entropy (see Section 4.3). The output of such a layer can be thought of as the *a-posteriori* probabilities $p(c|X)$ of the classes. Based on the assumption that both modalities used in this work are equally important and that the feature vectors extracted from the different modalities are statistically independent given the class, it is possible to apply the *sum rule*, probably the most widely applied approach for the late fusion of multiple classifiers, possibly corresponding to multiple modalities [27]:

$$\hat{c} = arg \max_{c \in \mathcal{C}} \{p(c|A) + p(c|S)\}, \quad (1)$$

where $\mathcal{C}$ is the set of all possible classes (*depression* and *control* in the experiments of this work), while A and S are the sequences extracted from the speech signal and its transcription, respectively (see section 4.1).

4.2.2 Feature Fusion. Section 4.1.3 shows that the feature vector sequences extracted from the speech signal and its transcription are encoded through the use of unimodal Bi-LSTMs that learn a representation capable to take into account relationships between the vectors in the sequence, possibly accounting for temporal patterns in the data. The two vectors resulting from such a process are L2-normalized and then fused according to multiple strategies. The first, referred to as *Feed Forward Feature Fusion (FF-FF)* in the following, corresponds to concatenating the unimodal enconding and feeding them to a feedforward network with four hidden layers (128, 64, 32 and 16 neurons, respectively). The expected effect of the hidden layers is to embed the encodings in a new, multimodal space more suitable for discriminating between depression and control paticipants. In a similar way, the second fusion strategy, referred to as *Logistic Regression Feature Fusion* (FF-LR) works by feeding the concatenation of the unimodal encodings to a Logistic Regression function trained to maximize the classification accuracy.

In both cases above, the assumption is that both modalities are equally effective at discriminating between depressed and control participants. However, this is not necessarily the case and, therefore, the last feature fusion stragey makes use of a *Gated Multimodal Unit* (GMU) and it is referred to as *Feature Fusion with Attention Gate FF-ATT*, where the GMU is a processing block that weights the different modalities through a self-attention mechanism [6]. If h_a and h_s are the encodings of speech signal and its transcription, respectively, the fusion is performed through a non-linear transformation that works according to the following equations:

$$x_a = tanh(W_a h_a) \quad (2)$$

$$x_s = tanh(W_s.h_s) \quad (3)$$

$$z = \sigma(W_z.[h_a \oplus h_s]) \quad (4)$$

$$h = z * x_a + (1 - z) * x_s, \quad (5)$$

where W_a, W_s and W_z are learnable parameters and $\oplus$ is the concatenation operator. The values of z and $1 - z$ can be thought of as weights that account for the contribution of the different modalities to the final classification outcome.

4.3 Classification

All representations, whether unimodal or multimodal, are fed to a fully connected *softmax* layer that implements the following equation:

$$\hat{l} = \sigma(W h_T + b), \quad (6)$$

where σ is the softmax function, W is the weight matrix and b is a bias vector. Both W and b are learned through a training process aimed at the minimization of the cross-entropy between groundtruth and classification outcome [12]:

$$\mathcal{L}(\mathcal{X}) = -\frac{1}{N} \sum_{n=1}^{N} [l_n \log \sigma(\hat{l_n}) + (1 - l_n)\ \log(l - \sigma(\hat{l_n}))], \quad (7)$$

where $\mathcal{X}$ is the training set, N is the total number of samples in $\mathcal{X}$, l_n is the groundtruth of training sample n and $\hat{l}_n$ is the classification outcome for the same sample. The training takes place through back-propagation with the use of gradient clipping to alleviate the exploding gradient problem [32].

5 EXPERIMENTS AND RESULTS

The goal of the experiments is to show whether it is possile to detect depression with less than 10 seconds of speech and, if yes, to what extent. For this reason, the rest of this section shows the performance of the approaches presented in Section 4 over the whole corpus at disposition and, as a comparison, it shows how such a performance changes when taking into account only the first clauses of an interview.

5.1 Hyperparameter Setting

The dataset has been split into 5 disjoint subsets through a random process such that all clauses belonging to a given subject are always in the same subset. In this way, it is possible to apply a k-fold approach ($k = 5$) and to perform participant-independent experiments, meaning that the clauses belonging to a given participant never appear in both training and test set. Every time a fold has been used as a test set, the union of the remaining four has been split into training set (90% of the material) and validation set (10% of the material). This latter has been used to select the value of the hyper-parameters via cross-validation. The space

Approach	Level	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Unimodal Text	Clause	60.4 ± 0.003	56.1 ± 0.005	46.5 ± 0.007	51.0 ± 0.005	59.0 ± 0.003
Unimodal Text	Participant	74.1 ± 0.023	100.0 ± 0.000	47.4 ± 0.047	64.1 ± 0.045	73.7 ± 0.023
Unimodal Audio	Clause	70.0 ± 0.006	65.1 ± 0.008	65.0 ± 0.008	65.0 ± 0.007	69.0 ± 0.006
Unimodal Audio	Participant	73.0 ± 0.021	76.0 ± 0.031	66.3 ± 0.031	71.0 ± 0.024	73.0 ± 0.021
Multimodal LF	Clause	64.0 ± 0.004	60.0 ± 0.006	54.3 ± 0.008	57.0 ± 0.005	63.0 ± 0.004
Multimodal LF	Participant	83.0 ± 0.036	94.0 ± 0.032	69.4 ± 0.068	80.0 ± 0.049	83.0 ± 0.036
Multimodal FF-FF	Clause	64.0 ± 0.004	59.3 ± 0.006	55.2 ± 0.008	57.2 ± 0.005	62.7 ± 0.004
Multimodal FF-FF	Participant	83.0 ± 0.027	93.1 ± 0.030	71.0 ± 0.043	80.1 ± 0.034	83.0 ± 0.027
Multimodal FF-LR	Clause	68.0 ± 0.006	64.3 ± 0.008	60.0 ± 0.010	62.0 ± 0.008	67.0 ± 0.006
Multimodal FF-LR	Participant	78.4 ± 0.021	85.0 ± 0.029	68.3 ± 0.033	76.0 ± 0.025	78.2 ± 0.021
Multimodal FF-ATT	Clause	63.0 ± 0.004	58.1 ± 0.005	54.5 ± 0.010	56.2 ± 0.007	62.0 ± 0.004
Multimodal FF-ATT	Participant	83.5 ± 0.031	95.0 ± 0.025	70.3 ± 0.058	80.5 ± 0.042	83.2 ± 0.031

Table 2: The table is shown the performance of unimodal and multimodal approaches used in the experiments, at both claus and participant level. The values are reported in terms of the averages obtained over 30 repetitions of the experiments and their standard errors.

of the hyperparameters (initial learning rate α_0, number of training epochs T, batch size B , number of hidden neurons for Bi-LSTM U and maximum length of an input sequence L) was searched through Gaussian Process Optimization. The models were trained using the Adam optimizer [26].

For the unimodal approaches, according to a practice common in the literature, the initial learning rate has been progressively reduced over successive training epochs using the expression $\alpha = \alpha_0 \beta^{\phi/\delta}$ where β=0.96 is the decay rate, ϕ is the step and $\delta = 500$ is the number of decay steps. In the case of the text model, the highest validation accuracy was obtained for $\alpha_0 = 0.003$, $T = 80$, $B = 64$, $U = 128$ and $L = 10$. For the Word Embedding, the experiments have made use of *itwiki*, the pre-trained Italian Wikipedia2Vec model, which is based on a 100-dimensional embedding space. In the case of the audio model, the hyperparameter values leading to the highest validation accuracy were $\alpha_0 = 0.001$, $T = 80$, $B = 32$, $U = 128$ and $L = 40$.

For the multimodal approaches (FF-FF and FF-ATT), the hyperparameter values maximizing the validation accuracy are $\alpha_0 = 0.003$ and $B = 128$. For FF-FF, the number of neurons in the 4 layers of the network is 128, 64, 32 and 16 (te values have been set a-priori and not through cross-validation). For FF-ATT, the size of the hidden layer in the gate is 27.

5.2 Recognition Results

Table 2 shows the performance of the approaches presented in Section 4. Since the training process starts with a random initialization of the parameters, every experiment has been replicated 30 times and Table 2 includes average and *Standard Error* (SE) across the 30 repetitions. The small SE values suggest that the variance is low and, hence, the models are robust to changes in the initial parameter values. Therefore, the average values can be considered realistic estimates of the actual performance of the approaches. According to a binomial test, all performances are better than chance to a statistically significant extent ($p < 0.001$ in all cases).

When it comes to unimodal approaches, the audio-based classifier performs better than the text-based one at the clause level and, according to a two-tailed t-test, the difference is significant ($p < 0.05$). However, from an application point of view, the most important metrics are those at the participant level and, in this case, the difference between audio and text is not statistically significant. At the clause level, multimodal approaches perform roughly like unimodal ones, but when it comes to participant level, the difference with respect to the best unimodal approach is always statistically significant ($p < 0.05$ according to a binomial test) except in the case of FF-LR. One possible explanation is that the unimodal encoders (see Section 4) capture temporal patterns in their respective input data, but represent them in a space where the difference between depression and control participants does not emerge with sufficient clarity. In this respect, the multi-layer network used in FF-FF to embed the unimodal encodings in a space where there is more difference between depression and control participants appears to lead to higher person level accuracy. FF-ATT does not make use of the four layers network, but it still achieves the same person level accuracy as FF-FF. In this case, the probable explanation is that the GMU effectively identifies the modality more likely to carry information leading to the correct classification.

5.3 Recognition and Number of Clauses

The last section shows that the application of the majority vote allows one to achieve high participant level accuracy, especially when it comes to multimodal approaches. The main probable reason behind such a result is that the average number of clauses per participant is greater than 100 for both depression and control participants (see Section 1). Therefore, a limited accuracy at the clause level is sufficient to increase the probability of at least half of the clauses being classified correctly, the condition for a participant being assigned to

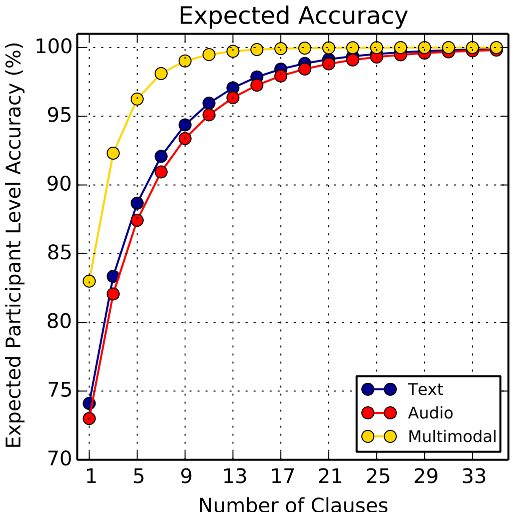

Figure 2: The plots show the expected accuracy of unimodal approaches and FF when using only a limited number of clauses. The expected accuracy is based on Equation (8) and it is based on the assumption that correctly classified clauses distribute uniformly across speakers.

the right class. In fact, such a probability can be estimated as follows (based on the assumption that the clause level accuracy is the same for all participants):

$$p_s = \sum_{k=M/2+1}^{M} \left(\begin{array}{c} M \\ k \end{array} \right) p^k (1-p)^{M-k}, \qquad (8)$$

where M is the average number of clauses per participant and p is the clause level accuracy. Figure 2 shows that such a probability increases significantly with the number of clauses and, therefore, the greater the number of these latter, the higher the expected participant level accuracy.

One of the main consequences of the considerations above is that it takes a substantial amount of time before the number of clauses is sufficiently large to ensure high performance. This is a problem for at least two reasons, namely the tendency of depressed people to speak less than the others (see Section 3) and the need to shorten the interviews in order to lower the costs associated to depression diagnosis. For this reason, this section investigates the relationship between performance and number of clauses. In particular, the analysis focuses on the two unimodal approaches and on FF-FF, the approach with the highest participant level recall.

Figure 3 shows how accuracy, precision and recall change as a function of the number of clauses used to make a participant level decision. The reason for taking into account only odd numbers is that this makes it possible to apply the majority vote without the risk of a tie. In terms of unimodal approaches, the plot shows that the accuracies of both audio and text unimodal approaches after one clause are within a statistical fluctuation with respect to the accuracies obtained while using the whole corpus. However, there are statistically significant differences for precision and recall. For both modalities, after the first clause, the precision is lower, but the recall is higher. For what concerns FF-FF, the pattern is similar, with the recall that has a small decrease (from 71.0% to 69.5%).

As the number of clauses increases, the pattern remains roughly the same for both unimodal and multimodal approaches. Therefore, the recall seems to improve or remain stable (in the case of FF-FF) when considering a limited amount of material. In this respect, using a limited number of clauses appears to ensure that more depression patients are recognised as such. Even if this comes at the cost of more control participants being classified as depressed, such a result can be considered positive because the consequences of type II errors (classifying a depression patient as control) are significantly more negative than those of type I ones (control participants classified as depression patients).

The effectiveness of the approaches after the first few clauses, especially in terms of recall, can lead to the interpretation that the depression patients tend to manifest their condition more clearly at the very beginning of the interview. Similarly, it can be argued that the results stem from the particular questions asked at the beginning of the interaction. For this reason, the same experiment has been conducted after shuffling the order of the clauses (see plots in the right column of Figure 3). It can be seen that the pattern is similar and this suggests that the clause order is not important. Furthermore, it confirms that using a limited amount of material appears to lead to the same recall level as when using the whole interview. Given that the average length of a clause is 1.2 seconds, the results above mean that such a time is sufficient to identify as many depression patients as those that get detected when using the whole material at disposition. In other words, it is possible to perform depression detection with less than 10 seconds without significant performance losses, especially when it comes to recall.

One possible explanation of the results above is that depression patients tend to manifest so consistently their condition, that there is high probability of correctly classifying any clause they utter. Not surprisingly, the clause level accuracy is well above chance for all approaches considered in the experiments. Such a result is in line with previous observations showing that limited amount of audio, possibly captured in naturalistic settings like the one of the experiments in this work, is sufficient to perform depression detection, especially when the approach is based on paralanguage [21]. On the other hand, the results of this work seem to contradict the finding in [1] that depression detection can improve by taking into account when a given sentence is uttered during a conversation.

6 CONCLUSIONS

This article has presented experiments aimed at showing whether it is possible to detect depression in less than 10 seconds and, if yes, how effectively. The experiments have been performed over a corpus of 59 clinical interviews and the

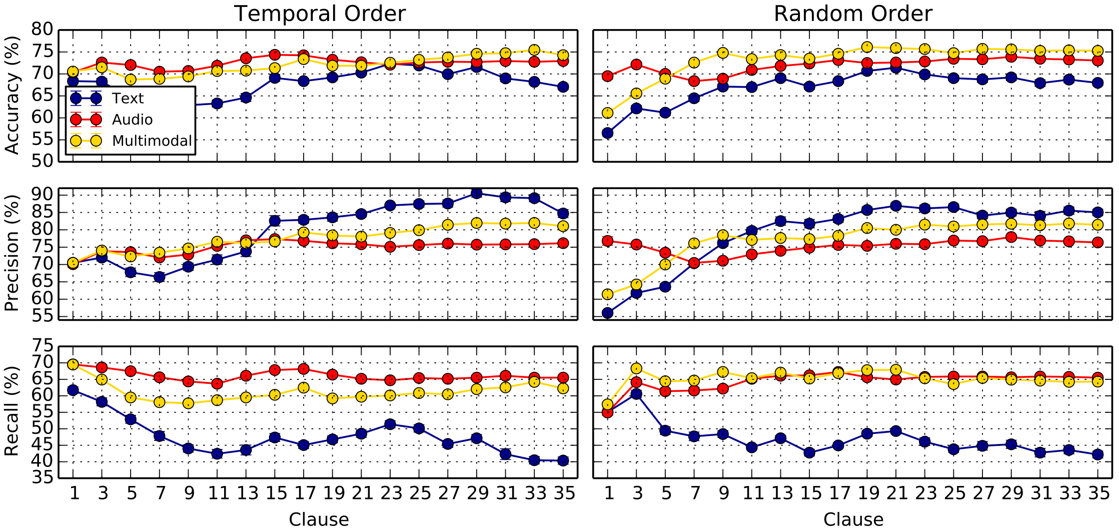

Figure 3: The plots show accuracy, precision and recall as a function of the number of clauses. The left column shows the results when the clauses are added in the same order as they appear in the interviews, while the right column shows the sam results when the clauses are added in random order.

results show that a few clauses - acatually accounting for less than 10 seconds - are sufficient to achieve a recall comparable (if not better) to the one obtained using the whole data at disposition. Furthermore, the results show that such a result can be observed whether the clauses are recognised in the order as they appear in the interviews or in a random order. This suggests that the observed results do not depend on the protocol applied at the beginning of the interviews, but on the amount of data.

The experiments have been performed using a wide spectrum of approaches aimed at the fusion of multiple modalities, including the combination of unimodal classifiers through the sum rule [27], one of the most traditional approaches for the combination of multiple classifiers, and network based approaches for the joint representation of multiple modalities [7], one of the most recent trends in multimodal behaviour analysis. Overall, multimodal approaches clearly outperform multimodal ones when using the whole corpus at disposition. However, this applies only to the participant level, after the application of the majority vote. When it comes to clause level accuracy, the performances of unimodal and multimodal approaches are actually closer and this probably explains why there are no major differences when taking into acccount one or a few clauses.

Fast depression detection addresses several issues in clinical practice. The first is the tendency of depressed individuals to avoid social interactions and to speak less than non-depressed people [9, 17]. The possibility to detect depression with limited material can help to deal with such a tendency and to obtain good results for people that cannot sustain an interview like those used in this work. The second is to spot actually depressed people among the many individuals that call counseling services because they are momentaneously in distress, but are not affected by a pathology. In this respect, approaches like those presented in this work can help to quickly dispatch callers among operators more or less qualified to deal with depressed individuals. Future work will focus on possible differences between depressed and non-depressed speakers in the modalities through which one's condition is manifested.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the project ANDROIDS funded by the program V:ALERE 2019 Università della Campania Luigi Vanvitelli, D.R. 906 del 4/10/2019, prot. n. 157264,17/10/2019. Alessandro Vinciarelli was supported by UKRI and EPSRC through grants EP/S02266X/1 and EP/N035305/1, respectively.

REFERENCES

[1] T. Al Hanai, M.M. Ghassemi, and J.R. Glass. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews.. In *Proceedings of Interspeech*. 1716–1720.
[2] M. Al Jazaery and G. Guo. 2019. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing (to appear)* (2019).
[3] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear. 2018. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing* 9, 4 (2018), 478–490.
[4] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear. 2013. Head pose and movement analysis as an indicator of depression. In *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction*. 283–288.
[5] L. Andrade, J.J. Caraveo-Anduaga, P. Berglund, R.V. Bijl, R. De Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller,

R.C. Kessler, N. Kawakami, C. Kiliç, D. Offord, T. Bedirhan Ustun, and H.-U. Wittchen. 2003. The epidemiology of major depressive episodes: Results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. *International Journal of Methods in Psychiatric Research* 12, 1 (2003), 3–21.

[6] J. Arevalo, T. Solorio, M. Montes-y G/'omez, and F.A. González. 2017. *Gated multimodal units for information fusion.* arXiv:1702.01992. arXiv.

[7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[8] A.T. Beck and B.A. Alford. 2009. *Depression: Causes and Treatment.* University of Pennsylvania Press.

[9] E.H. Bos, A.L. Bouhuys, E. Geerts, T.W.D.P. Van Os, and J. Ormel. 2006. Lack of association between conversation partners' nonverbal behavior predicts recurrence of depression, independently of personality. *Psychiatry Research* 142, 1 (2006), 79–88.

[10] H. Cai, X. Zhang, Y. Zhang, Z. Wang, and B. Hu. 2019. A case-based reasoning model for depression based on three-electrode EEG data. *IEEE Transactions on Affective Computing (to appear)* (2019).

[11] E. Charniak. 2018. *Introduction to Deep Learning.* MIT Press.

[12] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P.P. Kuksa. 2011. Natural Language Processing (almost) from Scratch. *CoRR* abs/1103.0398 (2011). http://arxiv.org/abs/1103.0398

[13] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski. 2015. Analysis of acoustic space variability in speech affected by depression. *Speech Communication* 75 (2015), 27–49.

[14] N. Cummins, V. Sethu, J. Epps, J.R. Williamson, T.F. Quatieri, and J. Krajewski. 2019. Generalized Two-Stage Rank Regression Framework for Depression Score Prediction from Speech. *IEEE Transactions on Affective Computing (to appear)* (2019).

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

[16] J. Garber, C.M. Gallerani, and S. A. Frankel. 2009. Depression in children. In *Depression in children*, I.H. Gotlib and C.L. Hammen (Eds.). The Guilford Press, 405–443.

[17] E. Geerts, N. Bouhuys, and R.H. Van den Hoofdakker. 1996. Nonverbal attunement between depressed patients and an interviewer predicts subsequent improvement. *Journal of Affective Disorders* 40, 1-2 (1996), 15–21.

[18] S. Gilbody, D. Richards, S. Brealey, and C. Hewitt. 2007. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine* 22, 11 (2007), 1596–1602.

[19] J.M. Girard, J.F. Cohn, M.H. Mahoor, S. Mavadati, and D.P. Rosenwald. 2013. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition.* 1–8.

[20] A. Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks.* Springer Verlag.

[21] Z. Huang, J. Epps, D. Joachim, and M. Chen. 2018. Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions.. In *Proceedings of Interspeech.* 3393–3397.

[22] C. Irons. 2014. *Depression.* Palgrave.

[23] J. Joshi, R. Goecke, G. Parker, and M. Breakspear. 2013. Can body expressions contribute to automatic depression analysis?. In *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).* 1–7.

[24] R.C. Kessler, P. Berglund, O. Demler, R. Jin, D. Koretz, K.R. Merikangas, A.J. Rush, E.E. Walters, and P.S. Wang. 2003. The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association* 289, 23 (2003), 3095–3105.

[25] R.C. Kessler and E.E. Walters. 1998. Epidemiology of DSM-III-R major depression and minor depression among adolescents and young adults in the national comorbidity survey. *Depression and Anxiety* 7, 1 (1998), 3–14.

[26] D.P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[27] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 3 (1998), 226–239.

[28] L. A. Low, N. C. Maddage, M. Lech, L.B. Sheeber, and N.B. Allen. 2011. Detection of Clinical Depression in Adolescents' Speech During Family Interactions. *IEEE Transactions on Biomedical Engineering* 58, 3 (2011), 574–586.

[29] F.A. McDougall, F.E. Matthews, K. Kvaal, M.E. Dewey, and C. Brayne. 2007. Prevalence and symptomatology of depression in older people living in institutions in England and Wales. *Age and Ageing* 36, 5 (2007), 562–568.

[30] M.R. Morales and R. Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *proceedings of the IEEE Spoken Language Technology Workshop.* 136–143.

[31] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing* 5, 3 (2014), 217–226.

[32] R. Pascanu, T. Mikolov, and Y. Bengio. 2012. Understanding the exploding gradient problem. *CoRR, abs/1211.5063* 2 (2012), 417.

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

[34] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. 2017. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the International Workshop on Audio/Visual Emotion Challenge.* 3–9.

[35] M. Rohanian, J. Hough, and M. Purver. 2019. Detecting Depression with Word-Level Multimodal Fusion. In *Proceedings of Interspeech.* 1443–1447.

[36] B. Schuller and A. Batliner. 2013. *Computational paralinguistics: emotion, affect and personality in speech and language processing.* John Wiley & Sons.

[37] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the International Workshop on Audio/Visual Emotion Challenge.* 3–10.

[38] M. Valstar, B. Schuller, J. Krajewski, J. Cohn, R. Cowie, and M. Pantic. 2014. AVEC 2014 – The Three Dimensional Affect and Depression Challenge. In *Proceedings of the ACM International Workshop on Audio/Visual Emotion Challenge.* 1–9.

[39] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. 2013. AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the ACM International Workshop on Audio/visual Emotion Challenge.* 3–10.

[40] J.C. Wakefield and S. Demazeux (Eds.). 2015. *Sadness Or Depression?: International Perspectives on the Depression Epidemic and Its Meaning.* Springer Verlag.

[41] WHO Document Production Services. 2017. *Depression and other common mental disorders.* Technical Report. World Health Organization.

[42] J.R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T.F. Quatieri. 2016. Detecting Depression Using Vocal, Facial and Semantic Communication Cues. In *Proceedings of the International Workshop on Audio/Visual Emotion Challenge.* 11–18.

[43] L. Yang, D. Jiang, and H. Sahli. 2019. Integrating Deep and Shallow Models for Multi-Modal Depression Analysis—Hybrid Architectures. *IEEE Transactions on Affective Computing (to appear)* (2019).

[44] Y. Yang, C. Fairbairn, and J.F. Cohn. 2012. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing* 4, 2 (2012), 142–150.

[45] X. Zhou, K. Jin, Y. Shang, and G. Guo. 2019. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing (to appear)* (2019).

[46] Y. Zhu, Y. Shang, Z. Shao, and G. Guo. 2017. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing* 9, 4 (2017), 578–584.