Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients

Alban Maxhuni [a], Angélica Muñoz-Meléndez [b], Venet Osmani [a,*], Humberto Perez [b], Oscar Mayora [a], Eduardo F. Morales [b]

[a] CREATE-NET, Via alla Cascata 56 D, Trento, Italy
[b] INAOE, Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla, Puebla, Mexico

ARTICLE INFO

Article history:
Received 12 January 2015
Received in revised form 15 September 2015
Accepted 8 January 2016
Available online xxxx

Keywords:
Motor activity
Bipolar disorder
Depression
Mental health
Speech emotions
Speech rate
Phone conversation
Personalized monitoring
Mobile computing

ABSTRACT

There is growing amount of scientific evidence that motor activity is the most consistent indicator of bipolar disorder. Motor activity includes several areas such as body movement, motor response time, level of psychomotor activity, and speech related motor activity. Studies of motor activity in bipolar disorder have typically used self-reported questionnaires with clinical observer-rated scales, which are therefore subjective and have often limited effectiveness. Motor activity information can be used to classify episode type in bipolar patients, which is highly relevant, since severe depression and manic states can result in mortality. This paper introduces a system able to classify the state of patients suffering from bipolar disorder using sensed information from smartphones. We collected audio, accelerometer and self-assessment data from five patients over a time-period of 12 weeks during their real-life activities. In this research we evaluated the performance of several classifiers, different sets of features and the role of the questionnaires for classifying bipolar disorder episodes. In particular, we have shown that it is possible to classify with high confidence (≈85%) the course of mood episodes or relapse in bipolar patients. To our knowledge, no research to date has focused on naturalistic observation of day-to-day phone conversation to classify impaired life functioning in individuals with bipolar disorder.

1. Introduction

The worldwide prevalence of many chronic health conditions is steadily increasing, so the management of diseases represents one of the most important challenges for health systems. The World Health Organization (WHO) has ranked mental disorders and mental injuries within the top 20 causes of disability among all medical conditions worldwide in persons aged in the range 14–55 [1]. Like other psychiatric disorder such as schizophrenia and major depression, bipolar disorder is a severe and chronic psychiatric illness that is associated with high rates of medical morbidity and premature mortality [2]. Illness characteristics and neurocognitive deficits certainly influence the quality of life and general functioning in bipolar disorder patients. One of its main characteristics is a repeated relapse of two polar episodes, mania and depression. Patients suffering from the disorder may experience episodes of altered mood states ranging from depression with sadness, hopelessness (including suicidal ideation), loss of energy, and psychomotor retardation, whereas manic episodes are

* Corresponding author.
E-mail addresses: alban.maxhuni@create-net.org (A. Maxhuni), munoz@inaoep.mx (A. Muñoz-Meléndez), venet.osmani@create-net.org (V. Osmani), humbertop@inaoep.mx (H. Perez), oscar.mayora@create-net.org (O. Mayora), emorales@inaoep.mx (E.F. Morales).

http://dx.doi.org/10.1016/j.pmcj.2016.01.008
1574-1192/

characterized by irritability, excessive energy (hyperactivity), reduction in the need of sleep and psychomotor agitation or acceleration.

The diagnosis of bipolar disorder is based on clinical evaluations through interviews and evaluations of scores gathered by quantitative psychopathological rating scales that were developed in the early 1960s (e.g. HAMD, BRAMS scales) and other more recent variations of them (e.g. BSDS). Although these interviews and questionnaires are well established and defined in a specific manual [3], they have their drawbacks, as they are performed on sporadic days, while a change to a potentially dangerous state can be produced in between these sessions. Other approaches include daily self-reports, however, they can be unreliable as they often depend on current mood episode polarity of the patients [4].

Currently, drug therapy is the main treatment, but its effectiveness critically depends on the timing of administration and has to be individually modified according to a patients' state of mind. Therapy can be very effective if administered at the beginning of a patient's transition to a different state but it may be less effective in severe states where the symptoms are present and persisted to a significant degree.

Motor activity is often used as a term to describe a group of symptoms that may range from mild to very severe, and is common feature of bipolar disorder. Assessing the motor activity of the patients with bipolar disorder has always been an essential part of psychiatric evaluations. Clinical measurement of motor activity is largely subjective and derives from caregivers' observations of specific behaviour. Motor functioning manifests itself in different areas such as speech production, facial expressions, gait, gestures, fine motor behaviour and the overall gross motor activity [5]. Furthermore, motor agitation has been shown to be potentially disruptive in patients with bipolar disorder who are experiencing a manic episode, a period when patients have increased activity levels, pressed to incoherent speech, racing thoughts and a decreased need for sleep. Motor activity may also be present during mixed and depressive episodes of bipolar patients, which can be reflected in motor retardation and irritable periods of time [3]. Therefore, monitoring motor activity is relevant for classifying critical state of the disorder. Smartphone is an enabling technology for this purpose due to increasing sensing capabilities.

The advantages of using technology to monitor bipolar disorder have recently been documented in the work carried out in MONARCA project [6–9] and have presented the basic concepts of using smartphones for the management of bipolar disorder. The authors emphasize the importance of state classification of patients by analysing several sensors embedded on a smartphone, such as location patterns from day-to-day activities, social-interaction sensing, level of physical activity, and phone usage and compared them with ground-truth values from psychiatric evaluations. Sensor data acquired from smartphones offers huge potential that through machine learning techniques get valuable insights of behaviour of bipolar disorder patients in their real life. In contrast to previous studies, we show that mood episodes of bipolar patients can be predicted using only information obtained during phone calls.

To our knowledge, no research to date has focused on a naturalistic observation of the day-to-day relationship between motor activities during phone conversation and patients' mood episode in individuals with bipolar disorder. This paper shows that motor activity features extracted from motion readings and speech articulation from smartphone sensors can be used to classify the course of mood episodes of a bipolar disorder patients. This is important because a non invasive and ubiquitous technology, like smartphones, can be used to obtain reliable information for patients during their phone conversations, in contrast to other studies using smartphone over long periods of time that can produce unreliable information when the phones are carried in purses, left at homes or use for playing or text messaging. In total, we analysed 2143 phone calls for a period of 12 weeks from five bipolar disorder patients. We extracted information from accelerometers, audio, and self-assessment questionnaires. We analysed the performance of several classifiers, the use of different sets of features, and the role of the questionnaires in the performance results. We achieved an average classification accuracy of 85.56%, and over 80% for all precision and recall values for each mental state of the different patients. This allows us to give fairly reliable information to clinicians that could take preventive actions on the effectiveness of medication and mitigate worsening of the condition.

The rest of paper is organized as follows: Section 2 summarizes previous research work with smartphones and concerning the monitoring of bipolar disorder patients. Section 3 provides information about the study group and how the data was collected. A detailed description of the attributes that were selected for classifying the episodes of the patients and representative graphs of their values on the training data is given in Section 4. Section 5 presents the experiments and results of our study, and discuss the potential benefits of the approach. Conclusions and future research directions are given in Section 6.

2. Related work

In this section we provide an overview of the most relevant research work on understanding human behaviour through mobile technology advancements.

2.1. Monitoring human behaviour through mobile technology

A number of researchers have demonstrated the potential of monitoring human behaviour using mobile computing and sensing technologies. Smartphones as ubiquitous devices are increasingly complex; they offer powerful computation and sensing capabilities for monitoring people's behaviour. They are often reported as deeply personal devices, regarding them as personal accessory [10]. Alongside these technological advances, there has been also increasing interest from researchers

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

and clinicians in harnessing smartphones as a means of delivering behavioural interventions for health. In the last decades there has been an increasingly wide range of research on using smartphone applications to support general mental and physical wellbeing. They have shown the effect of using smartphones in non-clinical settings for supporting regular physical activity and behavioural health, which is of critical importance for reducing the risks of several chronic diseases.

In recent years, there is a growing concern over the effects of sedentary behaviour lifestyle which have negative impacts on physical and mental health, such as increased risk of overweight, stress, fatigue and premature mortality [11]. In this line, several research work initiatives have suggested using smartphones for raising sedentary behaviour awareness and encourage regular physical activity by breaking their sedentary behaviour frequently during the day [12,13].

Emotional and social isolation are other aspects of users' behaviour that are indicative of overall wellbeing and mental health. Tools have been developed and used for sensing social engagements and quality of sleep in daily basis. For example, BeWell mobile applications monitor conversational turn-taking [14], while detection of stress from smartphone sensing has been described in [15–17].

Studies reviewed in this section show the potential of mobile computing with their embedded sensing capability, to effectively monitor patients' lives for a variety of health behaviours. As such, these systems have recently been used for monitoring patient's mental health, as discussed in the next section.

2.2. Mobile computing in bipolar disorder

Recent advances in mobile technology make smartphones increasingly relevant for monitoring different aspects of mental health care. Their affordability and sensing capabilities make them easier to use and able to provide clinicians better understanding of patients' behaviour, such as monitoring levels of physical activity, social interaction, sleep quality, mood and environmental factors, that could be used to enrich treatment assessments. Furthermore, mobile computing capabilities have created more opportunities for treatments to be available to patients during times and in situations when they are most needed, and to encourage activities between treatments sessions.

Several research work like [18,19] have used technological solutions to tackle mental health issues, which have shown significant potential for aiding emotional awareness. In this regard, some of the existing systems, such a Bliss Buzzer [20], include a continuous blood pressure monitoring tool for inferring the onset of stress and other behaviours related to stress. The authors demonstrate a functional system that collects and analyse physiological signals, and provide real-time feedback to end-users. To date, patients' assessments occur in clinical settings such as laboratory or hospitals, and the measurements rely on observation data collected in person, and often in laboratory setting. However, many aspects of mental diseases that are identified in clinical settings do not necessarily reflect what happens in real life activities. Therefore, several research work [21–23] suggest employing technological advancement for accurate and continuous monitoring of patients to reduce the impact of mental illness on a patient's daily activities and increasing the effectiveness of treatments.

Moreover, the usage of wearable and pervasive technology to infer mental functioning has been recently used in clinical studies. In this line, the research work in bipolar disorder have suggested combining physiological sensors together with smartphone sensing for monitoring different mental states of the patients. The work of [24] aimed at monitoring behavioural patterns using physiological data collected from smart textile platforms and smartphones. The objective of the study was measuring heart rate variability from respiratory rate aiming at predicting physiological changes that can be an indicator of bipolar disease. Each of these systems can quantify behavioural changes, but they require patients to wear sensors in a fixed location (e.g. ECG). However, wearing sensors can be a burden to the patients and may pose a discomfort in daily use caused by attached physiological sensors.

Other systems, like "Optimism App" [25] were designed to monitor depression and state changes, where patients are required to log self-reported mood, activities, physical activity level, and quality of sleep. Providing feedback about their health status to the patients, was recommended by psychiatrists for monitoring the patients' mental health. Similarly, the work of Burke et al. [26] suggests that mobile assessment provides more efficacious strategy for tracking mood and behaviour change in mental disorders than the traditional paper diaries.

However, despite their growth and performance for inferring behaviour in mental disorder diseases, very few solutions have been implemented in clinical practise for ambulatory monitoring of patients with bipolar disease. Several research work have shown how motor activity enhances mental health with respect to lowering levels of anxiety and depression, elevating mood, improving self-esteem and reducing stress [27,28].

As far as we know there is no reported study to date on information from daily motor activity taken during phone calls to classify the state of episode in patients with bipolar disorders.

2.3. Motor activity monitoring

It is well known that variations in levels of motor activity are an integral part in patients with bipolar disorder. Motor retardation has long been recognized as a feature of depressed patients, which differ from normal psychiatric comparison groups with regard to motor activity, body movements, speech and motor reaction time [29]. On the other hand, mania can be defined as motor agitation or restlessness with sustained fidgeting and frequent posture changes [30]. Therefore, motor activation may become a key factor in separating the states of the patients with bipolar disorder. Currently, the

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

clinicians assess measurement of motor activity during psychiatric hospitalization due to either mania or depression. Studies measuring level of motor activity in bipolar disorder have typically used traditional monitoring with paper and pencil diaries, and questionnaires [21]. These methods are often biased by the current state of the patient. Due to irritable state of individuals with bipolar disorder during depressive and manic state, patients are prone to neglect or to bias their reports on performed activities.

Monitoring motor activity during depression may be measured by actigraph [31]. In the actigraph, motor activity is measured using piezo-electric accelerometer that measures the intensity, amount and duration of movement in all directions. However, little is known if data captured from actigraph could provide motor activity characteristics in different states of bipolar disorder. Other studies using actigraph have reported the clinical findings that patients with depression are less active during daytime, and that activity increases during the treatment [32]. To our knowledge, studies using actigraph have not investigated difference in activity recorded by actigraph in patients having different episodes of bipolar disorder.

On the other hand, bipolar patients with remitted recurrent depression have been reported with frequent negative perception of social-interaction and motor retardation [33]. In this regard, smartphone could be a potential candidate for monitoring motor activity behaviour patterns in daily activities. Information from smartphones enables easier monitoring and tracking of the patients' progress than traditional methods, providing higher quality to the collected data. The benefits of using this technology include more accurate data from sensor-rich embedded on phones and also provide clinicians with the ability to evaluate the patients' progress in a more granular scale to increase the efficacy of the treatments.

In this paper, we collect data from accelerometers and speech characteristics during phone calls to infer motor activity changes in bipolar patients.

2.4. Monitoring characteristics of speech in bipolar disorder

One of the most accessible biological signals in human is speech, which can provide useful information about psychological conditions. An important area of research emerging in the field of psychiatry is the investigation of speech and emotion in bipolar patients in different states of the disease. Speech signal in bipolar disorder has been the focus of considerable research. Studies reported aspects of speech to be relevant in depression, such as spectral characteristics and content of speech articulation, the voice speed and loudness characteristics, and the verbal fluency produced during speech [34–37]. Several vocal parameters have been previously identified as possible cues to depression [34,37] and there is evidence, that these parameters can be used further to differentiate between bipolar disorder states.

Prosodic and spectral features were reported to vary in patients with regard to healthy individuals [34]. Vocal prosody as a composite of acoustic features of speech (i.e., semantic content of the signal) has been used in measuring and understanding the relation between the depression and prosody [35]. The main features used fundamental frequency (F_0) that could be perceived as pitch, intensity or loudness, speech rate, and rhythm. Moreover, glottal source parameter has shown relevance to pitch features and has been found to correlate with mood changes in bipolar patients and normal subjects [34].

Secondary features include jitter and shimmer, energy distribution among formants, and cepstral features. Many of these features have been explored with respect to emotion expression [34,37] and to a lesser extent to depression. Jitter has been shown to provide a good accuracy in short-term changes from the speech [34]. Thus, many studies have proposed jitter features as an important feature for the characterization of mood states, since it may reflect a deregulation of autonomous nervous system that influence muscular tone and articulator control [34,35]. On the other hand, shimmer, energy distribution among formants, and loudness feature groups have been reported to extract rapid variations in loudness from speech [34,36], such as irregularity of the vocal fold vibration, and lower energy for depressed patients caused by the glottal pulses. The above-mentioned studies have shown vocal features to correlate with depression. Although not completely consistent, they have been studied using automated classifiers for detecting depression. Despite their relative importance, their usage has been limited to laboratory settings without applications in bipolar disorder on clinical trials. A similar work to ours is the work from the University of Michigan that developed an smartphone application to detect an imminent mood swing in bipolar patients during phone conversations from speech characteristics [38]. Although the essential point of this work is similar to ours, we combine speech features and emotion features with motor activities during phone conversation, providing richer context information during phone conversations and we are able to identify differences in state moods from bipolar patients using information from speech features during phone calls during a 12 weeks period.

3. Trial description and data collection

In this section we describe the study group, the features considered in the tests and how the data was collected.

3.1. Participants and ground truth

The study group consisted of 10 patients (9 female and 1 male), all with a diagnosis of bipolar disorder. Patients have been categorized by the ICD-10, F31 (by the International Classification of Disease and Related Health Problems), with frequent changes of episodes. Recruitment of patients was performed according to inclusion criteria: age between 18 and 65, ability

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

Table 1
Number of calls and class associated to them based on psychiatric evaluations. There is also additional data (last column) where there is no class associated.

Patient	Severe depression	Moderate depression	Mild depression	Normal	Mild manic	Total	Additional data
P0201	36	–	113	149	–	298	435
P0302	135	–	–	99	–	234	199
P0702	–	112	39	–	–	151	116
P0902	–	142	–	161	–	303	178
P1002	–	–	35	–	162	197	28
All	171	254	187	409	162	1183	956

and are willing to operate modern smartphone devices. The patients were selected from the ward's psychiatrists that are capable of dealing with the requirements of the study. The trial was uncontrolled, not randomized, mono-centric, prolective, and observational study.

Psychiatric assessment and the psychological state examination were performed every 3 weeks over a period of 12-weeks at the psychiatric hospital Hall in Austria (TILAK—Department of Psychiatric, State Hospital, Hall in Tyrol, Innsbruck). The psychiatrists have set the interviews for the patients in such a way to reduce memory effect, which prevents having biased evaluation outcomes. To improve the scarcity of ground-truth, between scheduled interviews well trained and experienced clinicians talked collaboratively with patients about treatment by phone. During the examination, four standardized scales were used from clinical psychologists. Hamilton Depression Scale (HAMD) and Common Depression Scale (ADS) were used to determine depression, and Young Mania Rating Scale (YRMS) and Mania Self-Rating Scale (MSS) were used for determining mania. Evaluation of patients was normalized in a scale of -3 (severe depressed) to $+3$ (severe manic). Following the experiences in previous studies in MONARCA [6–9], also in the present study, ground-truth was applied using periods of 7 days before and 7 days after the examination with the basic assumption that the state changes are gradual.

3.2. Data collection

For this trial we used a Samsung Galaxy S2 device, although the software was developed to work with any compatible Android based phone. Patients were given a personal smartphone where the MONARCA application was set to start automatically on the phone startup and ran continuously in the background. The smartphone had a continuous sensing app installed that recorded on the phone memory and copied the information during periodic examinations. Raw data collected by the sensors included audio information during phone conversations, accelerometer, magnetometer, GPS, Wi-Fi Access Points, Bluetooth, phone usage, number of phone calls, SMS including their duration and length, and daily self-assessment questionnaires. The smartphone survey questions were designed to be equivalent to 22 questions of the HAMD and YMRS rating scales. The questions were selected to give an appropriate range of mood symptoms. Participants were required to respond in a scale (1–5) indicating the degree of their symptoms.

There were no constraints of any kind placed upon the patients, with respect to holding the phone in a specific manner or at a specific place in the body. Considering the fact that the trials were conducted under uncontrolled conditions in real life activities, in this study we focus on analysing accelerometer raw data and the speech features extracted from microphone during the phone conversation, when we are almost sure that the patients are holding their smartphone. We believe that both sensing techniques have their own advantages, complement each other, and can provide adequate information for classifying the course of mood episodes or relapse of a patient. In our experiments, we also included information from the self-assessment questionnaires relevant to motor activity, such as self-reported psychological state, physical state, and activity level.

We analysed the information collected and selected those patients with enough data recorded during their phone conversation and who represent different severities of disease on their psychiatric evaluation scores. Four of the patients refused to use the smartphone to make phone calls and one of the patients have very few calls. At the end we used information from five patients. This study also highlights the difficulties in obtaining real-world patient data, from the challenge in recruiting suitable patients up to patient compliance issues.

Table 1 shows the number of class and class associated to them based on psychiatric evaluations. It can be seen that we have a different number of calls per patient and per episode. The table also shows additional data (last column) indicating the numbers of phone calls that we have that are not associated to any episode as they were performed outside the 7 days window of the psychiatric assessments.

4. Feature selection

Feature selection from smartphone sensory data is probably the most important factor to consider in order to improve the recognition performance of machine learning tools. In the following subsections, we describe the most representative techniques for extracting time and frequency domain features from accelerometer raw data and prosodic and energy features extracted from speech.

Table 2
Features selected for the accelerometer sensor signals.

Time domain	Frequency domain
(1) Magnitude (2) Signal magnitude area (3) Root-Mean-Square (RMS) (4) Variance sum (5) Curve length (6) Non linear energy (7–14) For the 3 axes: Variance, Mean, Max, Min, Std. Dev., Absolute, Median, and Range (15–20) Mean and Std. Dev. of *X*, *Y* and *Z* axis. For all 20 features, we obtained the Min, Max, and Mean	(1) FFT energy (2) FFT mean energy (3) FFT StdDev energy (4) Peak power (5) Peak DFT bin (6) Peak magnitude (7) Entropy (8) DFT (9) Freq. Dom. Entropy (10) Freq. Dom. Entropy with DFT For all: Min, Max, Mean
Total: **60**	Total: **30**

4.1. Accelerometer signal features in time-domain (TD)

In order to quantify motor activities from the smartphone, acceleration readings collected during conversation (including picking up the phone, starting and finishing the call, and replacing the phone into the holder) were used in our analyses. These periods during conversation determine meaningful changes of acceleration values. We captured 3-axial linear acceleration continuously at rates, which varied due to Android system operating conditions, such as system load and battery levels. In this study, the accelerometer signals were resampled at a fixed rate of 5 Hz. The accelerometer features proposed in this paper, shown in Table 2, are quite popular amongst practitioners in the field, and were used as the basis for identifying periods of activity. To reduce the effect of spikes and noise from the accelerometer signal, statistical metrics such as mean, variance and standard deviation of three axis are applied over a window of approximately 26 s (non-overlapping fixed length windows of $n = 128$ samples).

Other features included the root-mean-square (RMS) acceleration for the period of conversation, as an indication of the time-averaged power in the signal. The RMS of a signal x_i, y_i and z_i represents a sequence of $n = 128$ discrete values obtained using Eq. (1).

$$\boldsymbol{RMS} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}. \tag{1}$$

The RMS results demonstrate differences in the motor activity during the phone conversation. The lower the RMS value, the lower the motor response which is manifested in depressed patients, whereas patients in the manic phase show elevated levels, as shown in Fig. 1(ii).

Another suitable measure for phone activities is the normalized signal magnitude area (SMA) that was used as the basis for identifying periods of activity during phone conversations, where $x(t)$, $y(t)$, and $z(t)$ are the acceleration signals from each axis with respect to time t as denoted by Eq. (2).

$$\boldsymbol{SMA} = \frac{1}{t}\int_o^t |x(t)|dt + \int_o^t |y(t)|dt + \int_o^t |z(t)|dt. \tag{2}$$

An example using SMA is presented in Fig. 1(i) where changes of motor activity can be compared in two states of the disease, transition from mild depressive state to normal state. The graph includes number of phone calls in both states ($n = 140$).

Also, a feature like Signal Vector Magnitude (SVM) [39] has been used to measure the degree of activity intensity and velocity of phone movement during the phone conversation and was obtained using Eq. (3). In addition to SVM, we computed the Variance Sum [40], that using the equation shown (4), where n represents the window size and *avgSVM* the mean of the *SVM* of that window size:

$$\boldsymbol{SVM} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i^2 + y_i^2 + z_i^2} \tag{3}$$

$$\boldsymbol{varSum}(n) = (SVM(n) - avgSVM(n))^2 - \left(\frac{n}{n-1}\right) - 2SVM(n). \tag{4}$$

Furthermore, in order to capture abrupt changes of phone activity during the phone conversation we used Averaged Non-linear Energy feature and Curve Length (CL) [41] feature using Eqs. (5) and (6).

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

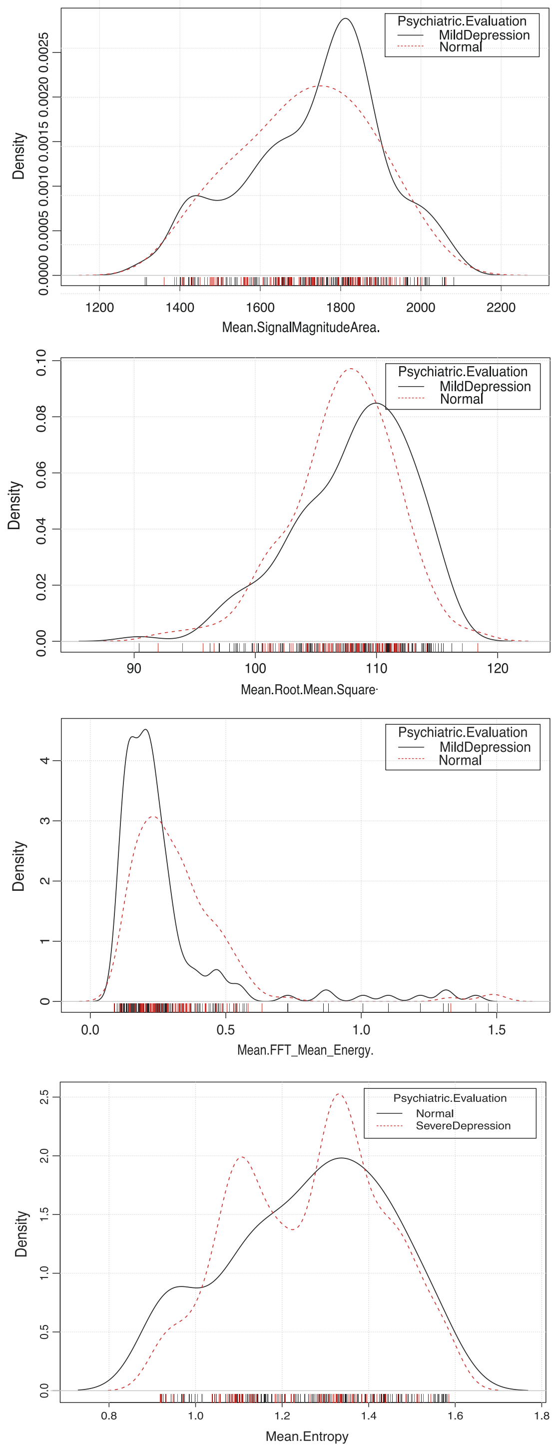

Fig. 1. Overall mean values of (i) RMS (p0201), (ii) SMA (p0201), (iii) energy (p0302) and (iv) entropy (p1002) with psychiatric evaluation.

$$\boldsymbol{CurveLength} = \sum_{i=1}^{n} |x_{i-1} - x_i| \tag{5}$$

$$NonE_i = x_i^2 - x_{(i-1)}x_{(i+1)}; \qquad avgNLE = \sum_{i=2}^{n-1} \frac{NonE_i}{n-2}. \tag{6}$$

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

4.2. Accelerometer signal features in frequency-domain (FD)

The signal and the distribution of signal energy over the frequency-domain are also popular choices in signal analysis. In this study, we used frequency-domain techniques to capture the repetitive nature of an accelerometer signal. These repetitions are often correlated to motor activity changes, which are capable of capturing distinctive pattern of movements in bipolar disorder patients during phone conversations. We applied the Fast Fourier Transform (FFT) on acceleration segments. Similarly as in TD, we used time window of approximately 26 s (non-overlapping fixed length windows of $n = 128$ samples), which enabled fast computation of FFTs that produces 128 components for each 128-sample window. Since our goal is to investigate the activity signatures, energy features were used to assess the strength of motor acts. The features in frequency-domain that are given in Table 2 have been used to determine the intensity of the signal. Total Energy of the acceleration signal was calculated as the squared sum of its spectral coefficients (sum of the squared discrete FFT component magnitudes of the signal) normalized by the length of the window. Using this metric, we were able to capture the intensity of the activity obtained using Eq. (7)component magnitudes of the signal.

$$energy = \sum_{j=1}^{(n/2)+1} y[j]^2. \tag{7}$$

Fig. 1(iii) shows an example of the total energy values of patient P0302 during phone conversations with different episode. As can be appreciated, the patient shows an increase level of motor activity in normal state compared to depressive states.

In order to determine the highest magnitude of all frequencies, frequency magnitude was measured using the real and imaginary components of the FFT values (using Eq. (8)). Frequency magnitude values below the cutoff and above the Nyquist rate ($\text{NyquistRate} = \text{windowLength}/2$) where nullified by keeping the peaks obtained in the window. Data has been normalized using Eq. (9) and multiplied by 2 to maintain the same energy. Furthermore, feature values obtained from entropy metric were measured using the normalized information entropy of the discrete FFT coefficient magnitudes by excluding the gravitational component, so called DC component of FFT (using Eq. (10)). Fig. 1(iv) shows an example of mean entropy values for patient P1002.

$$\textbf{Magnitude} = \sqrt{FFT.real^2 + FFT.imag^2} \tag{8}$$

$$\textbf{Normalized} = Magnitude * 2/windowLength \tag{9}$$

$$\textbf{Entropy} = \sum_{j=1}^{(n/2)+1} c_j \cdot log(c_j), \quad \text{where } c_j = \frac{|y_i|}{energy} \tag{10}$$

$$\textbf{Peak}_{Freq} =_j^{argmax} |y_i| \tag{11}$$

$$\textbf{Peak}_{energy} =_j^{max} |y_i|. \tag{12}$$

Together with the FFT Energy mean, FFT Energy standard deviation, FFT energy, DFT (Discrete Fourier Transform), and frequency magnitude, Entropy [42] is helpful in discriminating activities that differ in complexity. In our study, using this feature helped us to distinguish signals that have similar energy values with different motor activity patterns. Furthermore, we also investigated the largest signal peak using Peak Power Frequency that was compared against the baseline values (Eqs. (11) and (12)).

4.3. Feature selection and extraction from speech during the phone conversation

Previous work have shown scientific evidence that speech features can be used as an indicator of bipolar disorder [34,35]. In this regard, speech production is one physiological function that has been reported to affect motor retardation in bipolar patients. The application developed for our study, records speech signals from microphone only during the phone conversation with a sampling rate of 44 Hz and 16 bits amplitude quantization. Algorithms were developed to scrabbled/stretched the actual signal to avoid its original reconstruction while keeping the required properties for analysing the voice. In the current study, we extracted acoustic features from the speech signal using OpenEar [43] and Praat [44]. We evaluated features that have been successful in previous work [45]. Table 3 shows the acoustic features that were included in this research. We divided the features in two types: prosodic and vocal tract spectrum.

The features that were extracted from the patients speech data include the first-order functional of low-level descriptors (LLD) such as FFT-Spectrum, Mel-Spectrum, MFCC, Pitch, Energy, Spectral and LSP.39 functionals such as Extremes, Regression, Moments, percentiles, Crossings, Peaks, and Means. Prosodic features have been shown to provide rich source of information in speech such as pitch, loudness, speed, duration, pauses, and rhythm that could be used to detect the state of mind of patients during phone calls, i.e., when patients are in severe depressive state to normal or from moderate depression to normal states [34].

The second types of features were spectral features, which provide accurate distinction to a speaker's voice when prosodic aspects are excluded. We included the most popular voice quality descriptors shown in Table 3. With these types of features,

Table 3
Selected speech features relevant to bipolar disorder states.

Group	Feature type
Prosodic:	
Energy, Times	LOG energy, Zero crossing rate,
PoV, F_0	Probability of voicing, F_0
Spectral:	
MFCC	MFCC
MEL	MEL spectrum
SEB	Spectral energy in bands
SROP	Spectral roll of poing
SFlux	Spectral flux
SC	Spectral centroid
SpecMaxMin	Spectral max and min with DFT

Table 4
Selected speech features relevant to bipolar disorder states.

Emotional features	Spectral features
(1) % of Angriness	(1) No. of speech segments
(2) % of Nonconformity	(2) No. of short pauses
(3) % of Happiness	(3) No. of medium pauses
(4) % of Equanimity	(4) No. of long pauses
	(5) Total duration speech in call
	(6) Total duration not overlapped speech
	(7) Total duration overlapped speech
	(8) Quality of service
	(9) Duration of medium pauses in call
	(10) Duration of long pauses in call
Total: **4**	Total: **10**

Table 5
Overall number and duration of phone calls (incoming, outgoing) between the psychiatric assessments (Mean±SD).

Patient ID	1st–2nd PE	2nd–3rd PE	3rd–4th PE	4th–5th PE
0201	400 (8.76 ± 5.38)	204 (5.1 ± 3.7)	153 (4.02 ± 3.21)	193 (5.36 ± 3.79)
0302	169 (6.76 ± 3.4)	119 (5.66 ± 3.46)	158 (7.53 ± 4.52)	85 (5.31 ± 3.33)
0702	121 (6.1 ± 4.33)	50 (5.0 ± 3.01)	125 (7.73 ± 5.47)	119 (6.4 ± 4.92)
0902	172 (10.06 ± 7.09)	108 (8.71 ± 5.76)	185 (5.44 ± 4.85)	–
1002	130 (13.16 ± 8.01)	216 (11.36 ± 12.6)	–	–

we were able to distinguish periods of speech from patients, such as duration of speech segments, number and type of pauses (i.e., long, medium, and short), and overlapped or non-overlapped speech during conversations. We also measured the reaction and response time during the conversation time. We use the terms Number- and Duration of long pauses during the conversation to refer to the phone rate over the total conversation session, with times when the speech is not active (pauses) included in the total conversation session. Motivated by the clinical work carried out in studying bipolar patients in [35,18], we examined the association between long speech pauses in depressive patients and speech increments in manic phase during the phone conversation with their psychiatric scores, as shown in Tables 4 and 6.

Table 5 provides an overview of phone conversations during the trial. This table shows the overall number and average duration of phone conversation between the psychological evaluations in a daily basis. Since we focus on understanding meaningful information around the phone conversation, we keep accelerometer reading one-minute before the phone conversation, the readings from the entire duration of the call, and one minute after the conversation ended. Phone calls of less that 10 s were discarded in our experiments.

As can be seen from Table 6, average pauses and response delays in depressive state were inserted, in general, more often than during non-depressive state. This decrease can be seen across patients P0201, P0302, and P0902. In patients P0201 and P0302 it is more noticeable, where the average of decrease of phone call duration / average number of long pauses between the words went from 57.56(sec.)/0.52 during a depressive state to 39.77(sec.)/0.28 during a normal state (P0201); and patient P0302 where the average decrease of phone call duration and number of long pauses went from 130.86(sec.)/1.15 during a depressive state to 87.95(sec.)/0.87 during a normal state. In Fig. 2(iii) and (iv) we present the distribution of overall speech segments in conversation by mood episode of the patients. The speaking rate is significantly reduced during depressive periods as well as the duration of continuous speech segments.

In contrast to patients P0201 and P0302, where the transition of their state was from depression to normal phase, patient P0902 had a noticeable decrease number of long pauses during the phone calls as he went from a normal state to a depressive

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

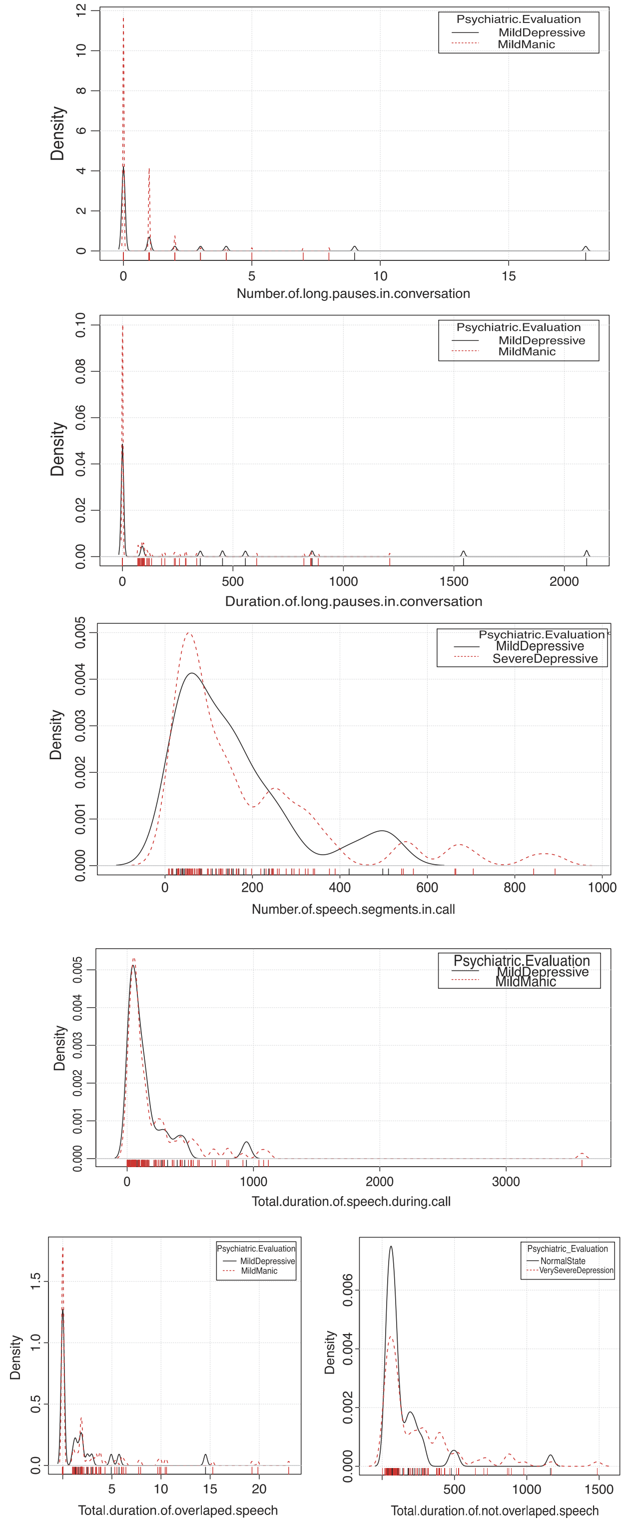

Fig. 2. Overall mean values of (i) number of long pauses (p1002), (ii) duration of long pauses (p1002), (iii) number of speech segments (p0702), (iv) number of speech segments (p0201), (v) duration of overlapped speech (p0302) and (vi) duration of not overlapped speech with psychiatric evaluation (P1002).

state. As such, there was a 26.91%/28.57% increase average duration of phone call duration and number of long pauses due to the transition to a depressive episode.

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

Table 6
Relationship between duration and number of long pauses in phone calls and psychiatric assessment scores (∗n/a-not applicable, since the patient did not experience a second depressive episode). Where MiD = Mild Depression, N = Normal, SeD = Severe Depression, MoD = Moderate Depression, and MiM = Mild Manic.

P.ID.	Avg. Duration / Avg. Long Pauses (Score)	Avg. Duration / Avg. Long Pauses (Score)	Difference (%)	Avg. Duration / Avg. Long Pauses (Score)	Difference (%)
P0201	57.56/ 0.52 (MiD)	39.77/ 0.28 (N)	**−30.90/ −46.15**	74.74/ 0.57 (SeD)	**87.93/ 103.57**
P0302	130.86/ 1.15 (SeD)	87.95/ 0.87 (N)	**−32.79/ −24.34**	n/a	n/a
P0702	54.19/ 0.53 (MoD)	143.66/ 1.32 (N)	**165.10/ 149.05**	119.17/ 0.98 (MoD)	**−17.04/ −25.75**
P0902	95.64/ 0.75 (N)	130.86/ 1.05 (SeD)	**26.91/ 28.57**	n/a	n/a
P1002	85.96/ 0.62 (MiM)	222.97/ 1.51 (MiD)	**73.27/ 143.54**	n/a	n/a

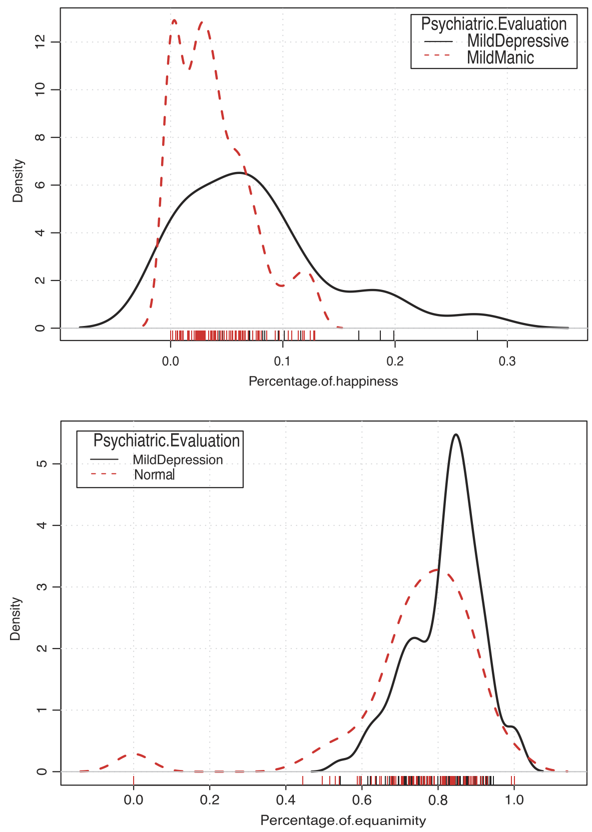

Fig. 3. Distribution of percentage of (i) happiness (p1002) and (ii) equanimity (p0201) features by psychiatric evaluation.

For the patient that experienced a manic episode, P1002 we can see a reverse trend, where the patient had decrease his average of long pauses, in accordance with the study reported in [37]. Average duration and number of long pauses were increased to 73.27%/143.54% during the depressive episode. Fig. 2(i) and (ii) provide the proportion of number/duration of long pauses between transitions from a manic episode to a depressive episode (P1002).

We also studied speech overlapping, voice quality and emotional features during phone conversations. Voice quality measures active speech frames, which were determined according to an energy-based speech activity. We explored the regularity and the responses from both active speakers during a phone conversation. Speech-overlapping was used to see the regularity during the conversation. Fig. 2(v) and (vi) present a comparison between non-overlapped in depression (P0302) and overlapped speech from patients in manic episodes (P1002).

The effects of emotional expression on speech are an interesting feature in bipolar disorder. Emotional state has been reported in previous studies, by identifying changes in muscle tension and in breathing. In our previous work [45], we have explored emotional state features from speech (i.e., happiness, angriness, nonconformity, and equanimity). In clinical reports that have investigated the symptoms in a manic episode, such as in [37], patients were characterized by extreme happiness and hyperactivity. Similarly, in our study we found a different percentage of happiness extracted from speech in manic episodes, while in a depressive state we found lower percentage of happiness, as shown in Fig. 3(i). Equanimity feature has also shown lower percentage in mild depression, whereas in normal state we found lower percentage of equanimity during the phone conversation (shown in Fig. 3(ii)).

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

Table 7
Number of features used in the experiments.

Feature	Number
Accelerometers:	
(1) Time-Domain	60
(2) Freq.-Domain	30
Audio:	
(1) Emotional	4
(2) Spectral	10
Questionnaire	3

Table 8
Accuracy results from different classifiers taken from Weka with their default parameters.

Classifier	P0201	P0302	P0702	P0902	P1002	Avg.(SD)	All P.
(a) Accuracy results from frequency domain features and all audio features.							
C4.5	89.93	85.47	**78.81**	87.79	85.79	85.56 (±4.17)	76.50
Random Forest (RF)	87.25	84.62	70.86	89.44	83.76	83.19 (±7.24)	70.33
SVM	**92.28**	75.21	75.50	83.50	**87.82**	82.86 (±7.52)	69.99
Naive Bayes (NB)	71.81	62.39	61.59	62.71	78.17	67.33 (±7.35)	47.59
k-NN (1)	87.90	63.68	59.60	79.54	81.22	74.39 (±12.14)	69.43
AdaBoost	84.56	**87.18**	74.17	**89.77**	86.80	84.50 (±6.06)	49.20
Bagging	89.26	86.32	71.52	89.44	86.29	**85.57** (±7.45)	**79.04**
(b) Accuracy results from frequency domain features and spectral features.							
C4.5	90.27	83.76	**78.81**	87.79	85.79	**85.28** (±4.35)	76.50
Random Forest (RF)	89.93	82.90	70.86	**90.43**	85.79	83.98 (±7.96)	**79.84**
SVM	**92.95**	75.21	76.16	83.83	**86.80**	82.99 (±7.44)	69.32
Naive Bayes (NB)	72.48	61.97	64.90	62.38	77.16	67.78 (±6.73)	46.83
k-NN (1)	87.58	63.38	61.59	77.89	81.73	74.43 (±11.46)	58.58
AdaBoost	84.56	**87.18**	74.17	89.77	87.82	84.7 (±6.17)	49.20
Bagging	89.26	86.32	72.85	89.77	86.29	84.90 (±6.93)	78.95

For our experiments we also used information from the questionnaires in terms of three attributes: (1) Physical, (2) Activity, and (3) Psychological, whose values range from 1 = low to 5 = high. A summary of all the attributes used in the experiments is given in Table 7. In the experiments we tested different sets of these attributes.

5. Experimental results

This section shows four experimental results in order to validate our model to classify bipolar disorder episodes with the available data:

1. Comparing the performance of different classifiers on the data
2. Selecting a set of features appropriate to the given task
3. Assess the effect of the information from the questionnaires on knowledge of depression in patients
4. Use a semi-supervised learning methodology to address the problem on how to use information from unlabelled data to enhance classification accuracy of bipolar disorder episodes from the phone calls information and specify the relationship between labelled and unlabelled data from entire data set.

We learned a model for each patient and also a single model combining all the information from all the patients. We performed 10-fold cross validation for all the experiments and report global accuracy, precision and recall values for each of the episodes.

5.1. Experiments with different classifiers

For our experiments, we used Weka's implementation of several classifiers with their default parameters. Table 8(a) and (b) shows the results from using emotional and spectral audio features with frequency domain features from the accelerometers and with information from the questionnaires. Similar results were obtained with other sets of features.

The table shows the accuracy results for different classifiers for each patient, their average, and the results for a single model with information from all patients (last column named "All P".). As can be seen from the table, there is no winning classifiers for all the data sets, although on average decision trees performed better than most other classifiers. It also performed reasonably well with information from all the patients. For these reasons and in the rest of the experiments we only report results for C4.5.

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

Table 9
Accuracy results from using different sets of features.

Features	P0201	P0302	P0702	P0902	P1002	Avg.(SD)
Accelerometer:						
– Time Domain (TD)	89.53	73.08	72.19	83.83	85.28	80.78 (±7.73)
– Frequency Domain (FD)	89.93	83.76	75.50	87.71	85.79	84.54 (±5.55)
Audio:						
– Emotional + Spectral	**90.60**	71.79	74.17	86.46	83.24	81.25 (±8.03)
– Spectral	**90.60**	73.93	74.17	87.12	83.24	81.81 (±7.55)
– Emotional	**90.60**	72.22	74.17	88.11	82.74	81.57 (±8.18)
– TD + Spectral	89.52	70.94	69.54	83.50	85.28	79.75 (±8.97)
– TD + Emotional	89.52	74.36	70.20	83.17	84.26	80.30 (±7.85)
– TD + (Emotional + Spectral)	89.53	70.09	69.54	82.18	85.28	79.32 (±9.07)
– TD + (Emotional + Spectral) without Questionnaire	50.0	59.40	70.86	51.16	81.22	62.53 (±13.37)
– FD + Spectral	90.26	83.76	**78.81**	87.78	85.79	85.28 (±4.34)
– FD + Emotional	90.27	**85.47**	74.83	87.79	**86.29**	84.93 (±5.93)
– FD + (Emotional + Spectral)	89.93	**85.47**	**78.81**	87.79	85.79	**85.56** (±4.18)
– FD + (Emotional + Spectral) without Questionnaire	79.53	84.19	74.83	**88.45**	85.79	82.56 (±5.40)

Table 10
Accuracy results using information from questionnaires. Results with an "*" indicate that the model is simply the majority class.

Patient	With quest.	Without quest.	Only quest.
P0201	89.93	79.53	90.60
P0302	85.47	84.19	76.92
P0702	78.81	74.83	74.17 (*)
P0902	87.78	88.45	89.77
P1002	85.79	85.79	82.74 (*)
Avg.:	85.55	82.55	82.84
All P.:	76.50	60.78	59.76

5.2. Different sets of features

We tested different sets of features. In particular, using only accelerometer features (time domain vs. frequency domain), only audio features (emotional and spectral), and combining accelerometer features with different audio features. In all these results information from the questionnaires was also included. Table 9 shows the results.

As can be seen from the experiments, using only features from the accelerometers have results over 80% on average with the frequency domain features performing slightly better. It is also interesting to notice that the audio features have similar performance, when both types, emotional and spectral, are considered together or when tested in isolation. The best results are obtained when the spectral and emotional features from audio are combined with the frequency domain features from the accelerometers.

For the rest of the experiments, all the results will be presented only with frequency domain features combined with the spectral and emotional features.

5.3. Impact from the questionnaire

Assessing the impact on the results from the questionnaires is important for producing a fully autonomous application. This is relevant as self-assessment can be counter-productive for depressed patients as they are reminded every day, with the questionnaires, of their state of depression. We performed tests with the frequency domain features and the audio features with and without information from the questionnaires, and also using only information from the questionnaires (examples with only three features). Table 10 shows the results.

As can be seen there is a small decrement in the results obtained without information from the questionnaires, however, the average results are still over 82%, from which it is reasonable to think in the development of a fully automatic monitoring tool. From the table it can be seen that using only information from the questionnaires, produces very competitive results. It is interesting to notice, that in this case, two models (marked with "*") are simply a majority class classifier, which of course are very poor classified for individual class values.

5.4. Semi-Supervised learning

As described in Table 1, there are more than 900 phone calls without an associated episode. In this subsection, we decided to use a semi-supervised algorithm to see if we can improve on the performance of previous results using all the available

Table 11
Accuracy results from a semi-supervised learning approach.

Patient	Supervised	Semi-supervised
P0201	81.54	83.78
P0302	85.47	81.53
P0702	72.84	71.45
P0902	87.45	88.77
P1002	85.76	83.78
Avg.:	82.61	81.86
All P.:	65.55	62.71

Table 12
Precision and recall results for some of the states of patients (MiD = Mild Depression, N = Normal, SeD = Severe Depression, MoD = Moderate Depression, and MiM = Mild Manic).

Patient	(State)	Precision		Recall	
		+Quest.	−Quest.	+Quest.	−Quest.
P0201	(MiD)	1.000	0.851	0.947	0.912
	(N)	0.890	0.826	0.919	0.799
	(SeD)	0.640	0.455	0.667	0.795
P0302	(SeD)	0.863	0.855	0.889	0.874
	(N)	0.842	0.823	0.808	0.798
P0702	(MoD)	0.851	0.843	0.866	0.813
	(MiD)	0.595	0.512	0.564	0.564
P0902	(N)	0.892	0.899	0.876	0.882
	(MoD)	0.862	0.869	0.880	0.887
P1002	(MiM)	0.899	0.904	0.932	0.926
	(MoD)	0.613	0.904	0.514	0.543
All P.:					
	– (SeD)	0.725	0.649	0.447	0.415
	– (MoD)	0.699	0.760	0.641	0.681
	– (MiD)	0.790	0.684	0.640	0.588
	– (N)	0.790	0.836	0.633	0.633
	– (MiM)	0.824	0.809	0.613	0.654
Avg.:		0.766	0.765	0.606	0.608

data. We followed a simple approach where we divided the data into ten folds; the training data was used to classify the unlabelled data. This classification included a weight associated with the classified value. We then used all the classified data with the original training set to produce an extended training set. We created a model with this set and test it on the testing set and then we averaged the results over the 10 folds. The results are presented in Table 11.

As can be seen from the results, adding information from other calls is not making much difference in the final results. There is a large number of alternative semi-supervised algorithms that can be considered in the future to improve over these results.

5.5. Precision and recall

Although the overall accuracy results may look promising, it is important to analyse the individual precision and recall measurements to see how effective the constructive models are for each episode. Table 12 shows the results for each patient for Mild Depression (MiD), Moderate Depression (MoD), Severe Depression (SeD), Mild Manic (MiM), and Normal (N) state.

It is interesting to note that most of the cases and both measures we have results above 80% with information from questionnaires and very similar results without information from questionnaires. We believe that these results give evidence that smartphone technologies can be effectively used as aid in the diagnosis of bipolar disorder episodes.

5.6. Predictive classes vs. expert evaluation

The last set of experiments was designed to show how the inductive models to classify the prospective onset of episodes of patients for all the available phone calls. We show only figures for the best results obtained with patient P0201 (Fig. 4) and for the worst results obtained with patient P0702 (Fig. 5). Both figures show at the top the evaluation scores from the psychiatrist, in the middle the classified states from the model, and the bottom the weight or confidence in the class classified by the model.

As can be seen from the figures, the induced models follow closely the assessment of the experts (which is not surprising as the models were trained with this information), and make reasonable classifications in the intermediate states between

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

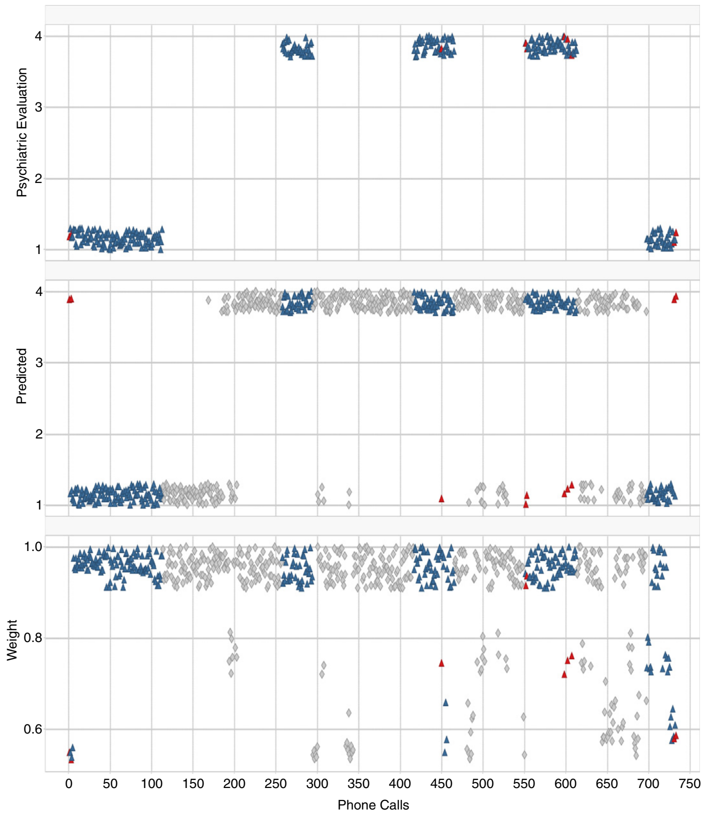

Fig. 4. Results from the induced model of patient P0201 and the assessments from the psychiatrist.

psychiatric assessments. The figures also show, in red, the classification errors produced by the models. In particular, it can be seen that the models make few errors, which can be further reduced, if the classification with a weight less than a threshold value, e.g., 0.8, are discarded.

6. Conclusions

In this paper we have presented a study of how to classify the course of mood episodes of bipolar disorder patients from information extracted from smartphones during phone calls. We used information from patients during 12 weeks on unconstrained conditions. We considered a wide range of features, both from accelerometer information and from audio information during the phone calls and analyse their behaviour for different users and mood episodes. We also make a comparison of different classifiers and different sets of features.

At least in this study, the information obtained from the frequency domain features of the accelerometers lead to higher classification accuracy than the information extracted from audio. Also, the frequency domain features produced better classification results than the time domain features. When we combined the audio features with the accelerometer features, there was only a small improvement when the emotional and spectral features were included.

Adding information from the questionnaires improved the overall results and also showed good results when considered on their own. However, without information from the questionnaires we obtained reasonable results (>80% for accuracy, precision and recall), suitable for the development of automatic tools that could aid psychiatrists in the monitoring of their patients.

As reported in other studies, personalized models behave better than single models that combined information from all patients. A future line of research is to construct prototype models using information from more patients, during longer periods of time, and with variations across different mood episodes.

We are considering as ground truth the doctor's assessment for a period of 14 days. It would be interesting to weight the doctor's assessment with respect to the number of days that occur before and after the consultation.

An obvious consequence of a good inductive model is to develop an application to alert doctors about possible warning signs or relapse of their patients. This could be useful to follow up on the effectiveness of medication treatments and it is critical to perform preventive measures on patients with severe depressed or manic states.

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008

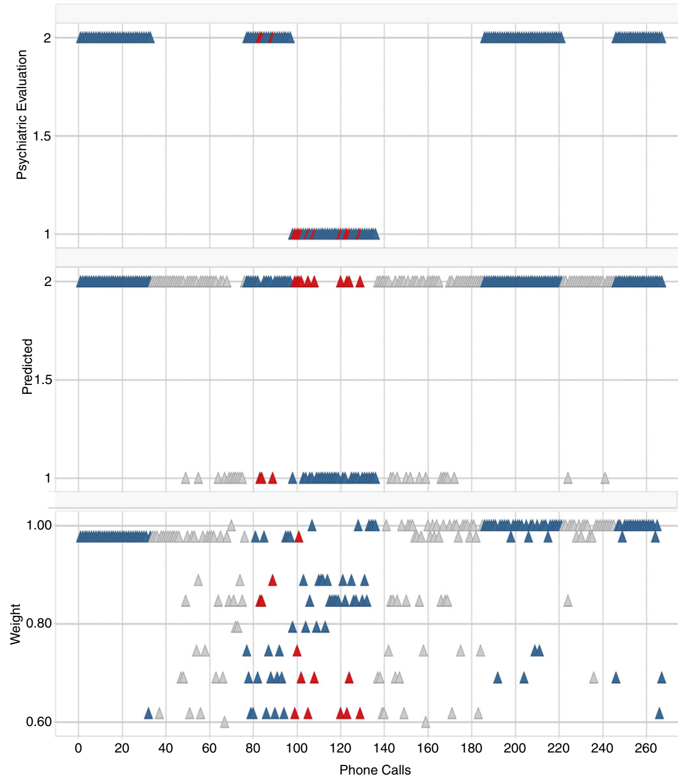

Fig. 5. Results from the induced model of patient P0702 and the assessments from the psychiatrist.

Acknowledgements

Authors acknowledge contribution of Embedded Intelligence Lab, DFKI, Germany and TILAK Clinic, Austria in data collection for this study. This work was supported by the EU FP7 MONARCA project (248545) http://www.monarca-project.eu.

References

[1] WHO. World health organization. the who world health report 2001: New understanding- new hope, world health organization. geneva. switzerland. 2001.
[2] J.M. Bopp, D.J. Miklowitz, G.M. Goodwin, W. Stevens, J.M. Rendell, J.R. Geddes, The longitudinal course of bipolar disorder as revealed through weekly text messaging: a feasibility study, Bipolar Disord. 12 (3) (2010) 327–334.
[3] M. Faurholt-Jepsen, S. Brage, M. Vinberg, E.M. Christensen, U. Knorr, H.M. Jensen, L.V. Kessing, Differences in psychomotor activity in patients suffering from unipolar and bipolar affective disorder in the remitted or mild/moderate depressive state, J. Affect. Disord. 141 (2) (2012) 457–463.
[4] J. Sims, F. Smith, A. Duffy, S. Hilton, The vagaries of self-reports of physical activity: a problem revisited and addressed in a study of exercise promotion in the over 65 s in general practice, Fam. Pract. 16 (2) (1999) 152–157.
[5] B.S. Alderfer, M.H. Allen, Treatment of agitation in bipolar disorder across the life cycle, J. Clin. Psychiatry 64 (2002) 3–9.
[6] V. Osmani, Smartphones in mental health: detecting depressive and manic episodes, IEEE Pervasive Comput. 14 (3) (2015) 10–13.
[7] A. Gruenerbl, V. Osmani, G. Bahle, J.C. Carrasco, S. Oehler, O. Mayora, C. Haring, P. Lukowicz, Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients, in: Proceedings of the 5th Augmented Human International Conference, ACM, 2014, p. 38.
[8] A. Grunerbl, A. Muaremi, V. Osmani, G. Bahle, S. Ohler, G. Troster, O. Mayora, C. Haring, P. Lukowicz, Smartphone-based recognition of states and state changes in bipolar disorder patients, Biomedical and Health Informatics, IEEE Journal of 19 (1) (2015) 140–148.
[9] V. Osmani, A. Maxhuni, A. Grünerbl, P. Lukowicz, C. Haring, O. Mayora, Monitoring activity of patients with bipolar disorder using smart phones, in: Proceedings of International Conference on Advances in Mobile Computing and Multimedia, MoMM'13, ACM, New York, NY, USA, 2013, pp. 85:85–85:92.
[10] L. Venta, M. Isomursu, A. Ahtinen, S. Ramiah, "my phone is a part of my soul"—how people bond with their mobile phones, in: Mobile Ubiquitous Computing, Systems, Services and Technologies, 2008. UBICOMM'08. The Second International Conference on, IEEE, 2008, pp. 311–317.
[11] S.J. Marshall, E. Ramirez, Reducing sedentary behavior a new paradigm in physical activity promotion, Am. J. Lifestyle Med. 5 (6) (2011) 518–530.
[12] S. Dantzig, G. Geleijnse, A.T. Halteren, Toward a persuasive mobile application to reduce sedentary behavior, Pers. Ubiquitous Comput. 17 (6) (2013) 1237–1246.
[13] H. Mukhtar, D. Belaid, Using adaptive feedback for promoting awareness about physical activeness in adults, in: Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), IEEE, 2013, pp. 638–643.
[14] N.D. Lane, M. Mohammod, M. Lin, X. Yang, H. Lu, S. Ali, A. Doryab, E. Berke, T. Choudhury, A. Campbell, Bewell: A smartphone application to monitor, model and promote wellbeing, in: 5th International ICST Conference on Pervasive Computing Technologies for Healthcare, 2011, pp. 23–26.
[15] E. Ceja, V. Osmani, O. Mayora, Automatic stress detection in working environments from smartphones' accelerometer data: A first step, IEEE J. Biomed. Health Inform. PP (99) (2015) 1–1.

[16] V. Osmani, R. Ferdous, O. Mayora, Smartphone app usage as a predictor of perceived stress levels at workplace, in: 9th International Conference on Pervasive Computing Technologies for Healthcare, IEEE, 2015.
[17] R. Ferdous, V. Osmani, J.B. Marquez, O. Mayora, Investigating correlation between verbal interactions and perceived stress, in: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, IEEE, Milano, 2015.
[18] C. Naranjo, C. Kornreich, S. Campanella, X. Noël, Y. Vandriette, B. Gillain, X. De Longueville, B. Delatte, P. Verbanck, E. Constant, Major depression is associated with impaired processing of emotion in music as well as in facial and vocal stimuli, J. Affect. Disord. 128 (3) (2011) 243–251.
[19] O. Schleusing, P. Renevey, M. Bertschi, S. Dasen, J.-M. Koller, R. Paradiso, Monitoring physiological and behavioral signals to detect mood changes of bipolar patients, in: Medical Information & Communication Technology (ISMICT), 2011 5th International Symposium on, IEEE, 2011, pp. 130–134.
[20] G. Quer, A. Nwokafor, A. Ganguly, N. Rashid, J. Zhu, D. Navani, R.R. Rao, Bliss buzzer, a system to monitor health and stress with real-time feedback, in: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, ACM, 2013, p. 83.
[21] M. Bauer, T. Wilson, K. Neuhaus, J. Sasse, A. Pfennig, U. Lewitzka, P. Grof, T. Glenn, N. Rasgon, T. Bschor, et al., Self-reporting software for bipolar disorder: validation of chronorecord by patients with mania, Psychiatry Res. 159 (3) (2008) 359–366.
[22] J. Blum, E. Magill, M-psychiatry: Sensor networks for psychiatric health monitoring, in: Proceedings of the 9th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNET 2008), 2008, pp. 33–37.
[23] R. Morriss, The early warning symptom intervention for patients with bipolar affective disorder, Adv. Psychiatr. Treat. 10 (1) (2004) 18–26.
[24] R. Paradiso, A. Bianchi, K. Lau, E. Scilingo, Psyche: personalised monitoring systems for care in mental health, in: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, IEEE, 2010, pp. 3602–3605.
[25] O.A.P. Ltd, Optimism apps. www.findingoptimism.com. Accessed: 26-Mar-2014.
[26] L.E. Burke, M.A. Styn, S.M. Sereika, M.B. Conroy, L. Ye, K. Glanz, M.A. Sevick, L.J. Ewing, Using mhealth technology to enhance self-monitoring for weight loss: a randomized trial, Am. J. Prev. Med. 43 (1) (2012) 20–26.
[27] J.B. Bartholomew, D. Morrison, J.T. Ciccolo, et al., Effects of acute exercise on mood and well-being in patients with major depressive disorder, Med. Sci. Sports Exerc. 37 (12) (2005) 2032.
[28] A.L. Dunn, M.H. Trivedi, J.B. Kampert, C.G. Clark, H.O. Chambliss, Exercise treatment for depression: efficacy and dose response, Am. J. Prev. Med. 28 (1) (2005) 1–8.
[29] D. Todder, S. Caliskan, B.T. Baune, Longitudinal changes of day-time and night-time gross motor activity in clinical responders and non-responders of major depression, World J. Biol. Psychiatry 10 (4) (2009) 276–284.
[30] L.L. Judd, P.J. Schettler, H. Akiskal, W. Coryell, J. Fawcett, J.G. Fiedorowicz, D.A. Solomon, M.B. Keller, Prevalence and clinical significance of subsyndromal manic symptoms, including irritability and psychomotor agitation, during bipolar major depressive episodes, J. Affect. Disord. 138 (3) (2012) 440–448.
[31] K.A. Kaplan, L.S. Talbot, J. Gruber, A.G. Harvey, Evaluating sleep in bipolar disorder: comparison between actigraphy, polysomnography, and sleep diary, Bipolar Disord. 14 (8) (2012) 870–879.
[32] C. Burton, B. McKinstry, A. Szentagotai Tătar, A. Serrano-Blanco, C. Pagliari, M. Wolters, Activity monitoring in patients with depression: A systematic review, J. Affect. Disord. 145 (1) (2013) 21–28.
[33] E.H. Bos, A.L. Bouhuys, E. Geerts, T.W. van Os, I.D. van der Spoel, W.H. Brouwer, J. Ormel, Cognitive, physiological, and personality correlates of recurrence of depression, J. Affect. Disord. 87 (2) (2005) 221–229.
[34] E. Moore, M. Clements, J. Peifer, L. Weisser, Analysis of prosodic variation in speech for clinical depression, in: Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE, vol. 3, IEEE, 2003, pp. 2925–2928.
[35] E. Moore, M.A. Clements, J.W. Peifer, L. Weisser, Critical analysis of the impact of glottal features in the classification of clinical depression in speech, IEEE Trans. Biomed. Eng. 55 (1) (2008) 96–107.
[36] A. Ozdas, R.G. Shiavi, S.E. Silverman, M.K. Silverman, D.M. Wilkes, Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk, IEEE Trans. Biomed. Eng. 51 (9) (2004) 1530–1540.
[37] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanata, E.P. Scilingo, Speech analysis for mood state characterization in bipolar patients, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, pp. 2104–2107.
[38] S. Aimee, It's not just what you say, it's how you say it. http://www.slate.com/articles/technology/future_tense/2014/08/priori_smartphone_app_listens_to_bipolar_patients_to_detect_manic_episodes.html. Accessed: August 3 2014.
[39] D.-U. Jeong, S.-J. Kim, W.-Y. Chung, Classification of posture and movement using a 3-axis accelerometer, in: Convergence Information Technology, 2007. International Conference on, IEEE, 2007, pp. 837–844.
[40] FUNF. Funf - open sensing framework. https://code.google.com/p/funf-opensensing-framework/source/browse/src/edu/mit/media/funf/probe/builtin/activityprobe.java?r=55f67cf53180dbba8ed4eddef422737f3abe030e Accessed: 2015-AUG-29.
[41] S. Mukhopadhyay, G. Ray, A new interpretation of nonlinear energy operator and its efficacy in spike detection, IEEE Trans. Biomed. Eng. 45 (2) (1998) 180–187.
[42] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons, 2012.
[43] F. Eyben, M. Wollmer, B. Schuller, Openear—introducing the munich open-source emotion and affect recognition toolkit, in: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, IEEE, 2009, pp. 1–6.
[44] P. Boersma, Praat, a system for doing phonetics by computer, Glot international 5 (9/10) (2002) 341–345.
[45] H. Pérez-Espinosa, C.A. Reyes-García, L. Villaseñor-Pineda, Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model, Biomed. Signal Process. Control 7 (1) (2012) 79–87.

Please cite this article in press as: A. Maxhuni, et al., Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.01.008