Assessing Bipolar Episodes Using Speech Cues Derived from Phone Calls

Amir Muaremi[1(✉)], Franz Gravenhorst[1], Agnes Grünerbl[2], Bert Arnrich[3], and Gerhard Tröster[1]

[1] Wearable Computing Lab, ETH Zurich, Gloriastrasse 35, 8092 Zurich, Switzerland
{muaremi,gravenhorst,troester}@ife.ee.ethz.ch
[2] TU Kaiserslautern, Embedded Intelligence, 67663 Kaiserslautern, Germany
agnes.gruenerbl@dfki.de
[3] Computer Engineering Department, Boğaziçi University, 34342 Istanbul, Turkey
bert.arnrich@boun.edu.tr

Abstract. In this work we show how phone call conversations can be used to objectively predict manic and depressive episodes of people suffering from bipolar disorder. In particular, we use phone call statistics, speaking parameters derived from phone conversations and emotional acoustic features to build and test user-specific classification models. Using the random forest classification method, we were able to predict the bipolar states with an average F1 score of 82 %. The most important variables for prediction were speaking length and phone call length, the HNR value, the number of short turns and the variance of pitch F_0.

Keywords: Bipolar disorder · Smartphone · Voice analysis · Phone calls

1 Introduction

1.1 Motivation

Bipolar disorder is a mental illness characterized by alternating episodes of mania and depression. About 2.4 % of people worldwide are diagnosed with bipolar disorder at some point in their lifetime; in the USA this figure reaches 4.4 % [11]. This illness is responsible for more handicapped life-years than all forms of cancer and one in four bipolar patients have a history of attempted suicide. Each chronic case often causes lifetime costs of more than $600,000 [2]. The state-of-the-art method for diagnosis and monitoring of bipolar disorder centers on frequent visits to the doctor and self-assessment questionnaires. These methods are time-consuming, expensive and rely on the availability of experienced doctors, making them particularly hard to implement in low-income countries [11]. We envisage supporting the diagnosis and monitoring of bipolar disorder patients

This work is sponsored by the FP7-ICT MONARCA project (ref. no. 248545).

P. Cipresso et al. (Eds.): MindCare 2014, LNICST 100, pp. 103–114, 2014.
DOI: 10.1007/978-3-319-11564-1_11

with technical means, particularly through the use of smartphones. This could potentially increase the affordability, availability and pervasiveness of treatment for patients. This is a realistic possibility considering the current trend towards decentralized pervasive healthcare [1] and the fact that there are over 4.5 billion unique mobile phone users, in other words more than 60 % of the world population owns at least one mobile phone, 30 % of which are smartphones [4]. For economic and usability reasons it makes sense to exploit these already existing hardware devices rather than developing new ones. In this work we explore the potential of smartphones for monitoring bipolar patients by focusing on voice analysis as the most natural modality available.

1.2 Related Work

Very recently, new approaches for monitoring and detecting mental disorders with the help of wearable and mobile devices have been investigated. Frost et al. [7] transferred the traditional and well-established methodology of paper-based questionnaires to wearable devices such as phones or tablets. Moore et al. [12] used short text messages to collect self-assessment data from 153 bipolar patients. This data is used to evaluate models to forecast manic and depressive episodes. The authors conclude that self-assessment is very heterogeneous and this constrains the accuracy of their forecast models. Besides self-assessment questionnaires, analysis of patients' voices is another well-established method for diagnosing affective disorders. Voice analysis studies noted in the literature date back to as early as 1938 [13]. In the area of neurodegenerative diseases speech analysis is very accurate. Applying speech analysis, Tsanas et al. [15] manage to discriminate Parkinson patients from healthy controls with a 99 % accuracy. For classifying mental disorders, psychiatrists usually follow well-established guidelines, assessment protocols or rating scales. Many of these state-of-the-art rating scales involve statements related to the patient's voice. For example, the Young Mania Rating Scale [18] requires the psychiatrist to assess the speech rate and how much the patient talks. Vanello et al. [16] uses signal processing algorithms to analyze voice automatically. The method is applied for episode identification of bipolar patients in a controlled environment.

To analyze speech in real-life situations an unobtrusive, wearable device with a microphone has to be carried by patients. Mobile phones fulfill these criteria and are carried anyway by a vast majority of the population. One of the first studies involving mobile phones that analyzed users' voices is presented by Lu et al. in 2012 [10]. The assessment of the users' stress levels in unconstrained outdoor acoustic environments achieved an accuracy of 76 %. Xu et al. [17] analyze voice data to estimate the number of speakers in a room. Other approaches exploit the internal sensors implemented in smartphones, like GPS positions, accelerometer data or Bluetooth fingerprints. These approaches use data mining to recognize activity patterns and classify mental states [8,9,14].

In this work we explore the feasibility of voice analysis during phone conversation with smartphone microphones to predict bipolar disorder episodes.

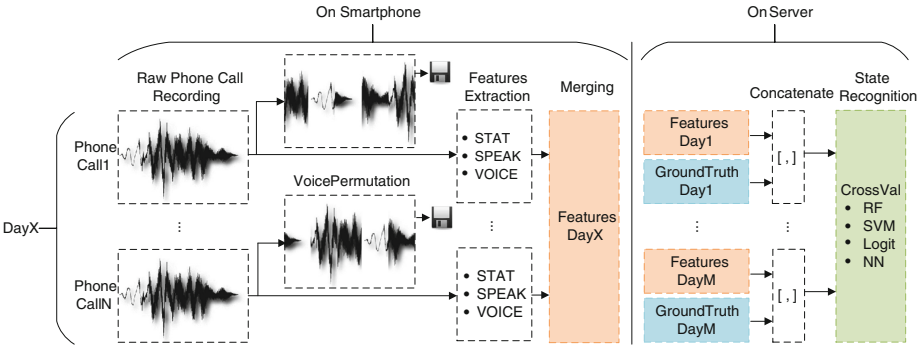

Fig. 1. System overview: from recording the voice to the state recognition

1.3 System Overview and Paper Organization

Figure 1 provides an overview of the chain from phone call recording to predicting bipolar states. During phone conversation, the speech of the patient is recorded. Immediately after the call ends, its features are extracted from the raw recording. Before the original file is deleted, a scrambled version of it is stored on the smartphone. At the end of the day, the features of all phone calls during that day are merged together, resulting in one data point per every 24 h. On server-side, daily features are concatenated with the corresponding ground-truth scores. These pairs build the input for training and testing different classifiers in a cross-validation manner in order to assess the prediction performance of the current state of a bipolar patient.

The rest of the paper is organized as follows: details about the data collection are presented; next, we describe three different phone call feature sets used for prediction analysis; Sect. 4 shows the classification performance and the parameters that best contribute to predicting bipolar episodes; and the work is concluded by discussing the limitations of the study and summarizing its main achievements.

2 Data Collection

In this section we briefly describe the data collection trial and how the ground-truth is derived. We also discuss the integrity of the collected data and show how participants' privacy is maintained (see [9] for details).

2.1 Trial Description

A data collection trial was deployed in cooperation with the psychiatric hospital Hall in Tirol, Austria. A total number of 12 bipolar patients between the age of 18 and 65 were recruited during a stationary stay at the psychiatric hospital. After signing an informed consent form, they were provided with an Android

smartphone and were asked to collect behavioral data over a time-span of 12+ weeks. Apart from automatically collecting smartphone data, the patients were asked to submit a daily subjective self-assessment by filling out a questionnaire. The trials were 'real-life', meaning that the patients were encouraged to use the smartphone as their normal smartphone, with no restrictions or limitations.

2.2 Handling of Ground-Truth

Objective ground-truth of the patients' state was gathered every three weeks at the hospital. Standardized psychological tests scales for depression and mania were used, combined with psychiatric assessments. These measurement points resulted in an assessment on a scale between −3 (heavily depressed) to +3 (heavily manic), with intermediate steps of depressed, slightly depressed, normal, slightly manic and manic. An overview of the assessment for each patient over the study duration is depicted in Fig. 2. In a normal case, the number of ground-truths is 5 (hospital visits) per patient, which is very few for analysis. The following procedure was applied to extend the number of ground-truth days: None of the patients were rapid-cyclers, i.e. the change of state did not happen within a few days but rather at least one or more weeks, yet changes of state would likely happen after a visit to the doctor (during the examination point). Therefore, according to experienced psychiatrists it was acceptable to project the ground-truth assessment values 7 days before the examination and 2 days

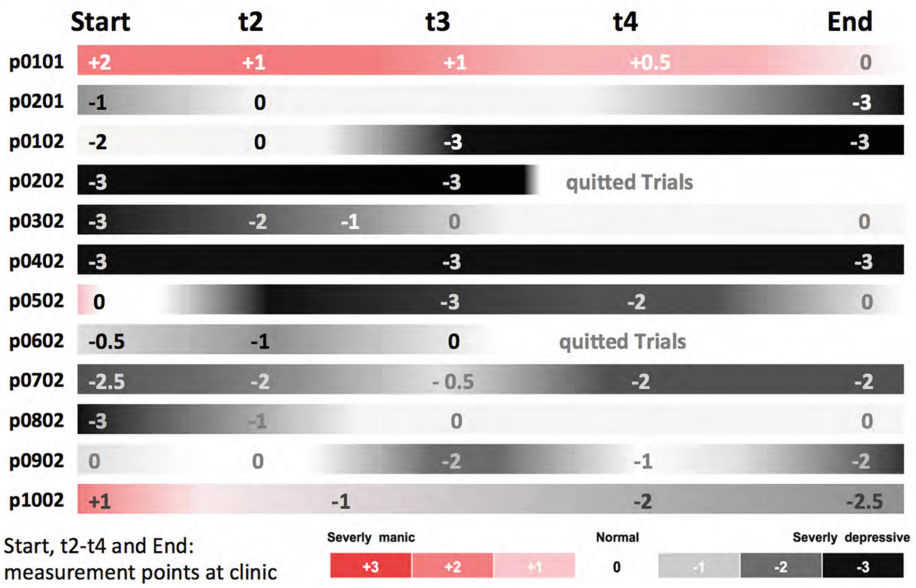

Fig. 2. States of the individual patients at different measurement points [9]

after. This time-period was adjusted (extended or shortened) according to stable or unstable daily subjective self-assessments.

2.3 Data Completeness

The maximal amount of data (5 measurement points for 12 patients times 9 days (7 before + 2 after)) was reduced due to some practical factors. Two patients (p101, p802) did not use the new smartphone for phone-calls but kept their old cell-phone for this. Furthermore, two patients (p202, p402) did not show any changes in state during the entire trial time. Therefore, their data was of no use in respect of state classification and had to be discarded. Moreover, the presence of ground-truth together with the availability of phone-call data was necessary, yet sometimes patients switched off their smartphone for several days or did not receive or conduct phone-calls and therefore, little or no data was available for the measurement point. As a consequence of these factors, only 6 of 12 patients (p201, p102, p302, p602, p902, p1002) provided enough data points for different mental states to make reliable classification possible (see Table 1). Each of these patients experienced only 2 out of 3 phases during the entire trial period (see Fig. 2), resulting in a two-class problem for the later state recognition.

Table 1. The number of ground truth (GT) days and voice data per patient. The last row presents the data distribution in the two GT classes per patient.

# of Days	Patients											
	p101	p201	p102	p202	p302	p402	p502	p602	p702	p802	p902	p1002
Total	97	83	75	??	90	??	131	53	76	115	91	67
GT	84	47	52	??	70	??	63	41	53	71	48	47
Voice	0	79	66	42	83	41	4	46	62	0	89	61
GT + Voice	0	37	44	0	58	0	0	33	36	0	41	42
Classes	-	12|25	31|13	-	17|41	-	-	12|21	32|4	-	26|15	11|31

2.4 Privacy Compliance

The main requirement of the ethical committee was to ensure that the semantic content of the stored speech on the smartphone was not accessible at any time. To ensure this, we cross-compiled the feature extraction code for ARM processors resulting in a toolbox, which can be run on android phones. Immediately after the phone call is finished this toolbox is used to derive the features directly on the smartphone. These high-level features (see next section), with which the speech cannot be reconstructed, are stored locally on the phone.

Before the speech recording is deleted, a modified version of the original file is created. Each 0.5 s segment of the speech is divided into 25 ms chunks. These 20 slices are randomly permuted and for each segment the permutation

order is changed. A low-pass filter is applied at the end to remove the signal discontinuities, i.e., jumps, at borders of the chunks and segments. The content of the resulting concatenated audio file is not understandable at all, i.e., the speech intelligibility is zero. The scrambled version of the audio file will be used for improving the bipolar state recognition algorithms in the future.

3 Phone Call Features

We differentiate between statistical features (`STAT`), speaking cues (`SPEAK`) used in social signal processing and voice features (`VOICE`) used in the area of acoustic emotion recognition. The calculation of the features is based on the open-source audio feature extractor "openSmile" [5]. In the following sections the categories are described in more detail.

3.1 Phone Call Statistics

The basic phone call statistics are derived from the meta data of the speech file without considering the content of that file. The following `STAT` features were calculated on a daily basis:

1. Number of phone calls during the day
2. Sum of the duration of all phone calls during the day
3. Average duration of the phone calls
4. Standard deviation of phone call durations
5. Minimum duration of all daily phone calls
6. Maximum duration of all daily phone calls
7. % of phone calls in the morning (between 4am and 9am)
8. % of phone calls in the night (between 11pm and 4am)

3.2 Speaking Cues

From the voice recordings we extract non-verbal activity cues adopted from [6] to describe the speaking behaviour of a patient in a conversation. Based on the output of voice activity detection (voiced speech vs. unvoiced speech) the speaking segments are created. Speaker diarization is not necessary since the audio recording contains only the voice of the patient. Figure 3 shows an exemplary audio recording and the highlighted speaking segments. In a conversation a speaker turn is the time interval when that person is speaking. Short turns or utterances are most likely to be back-channels, i.e., feedback words while someone else is talking, such as "okay", "hm", "right", etc. Non-speaking segments are either pauses or turns from the other person on the line. The following `SPEAK` features were calculated on a daily basis:

1. Average speaking length ($\texttt{STAT}_3$ without the non-speaking segments)
2. Average number of speaker turns

3. Average speaking turn duration
4. Standard deviation of speaking turn duration
5. Average number of short turns/utterances
6. % of speaking from the total conversation
7. Speaker turns per length in minutes
8. Short turns/utterances per length in minutes

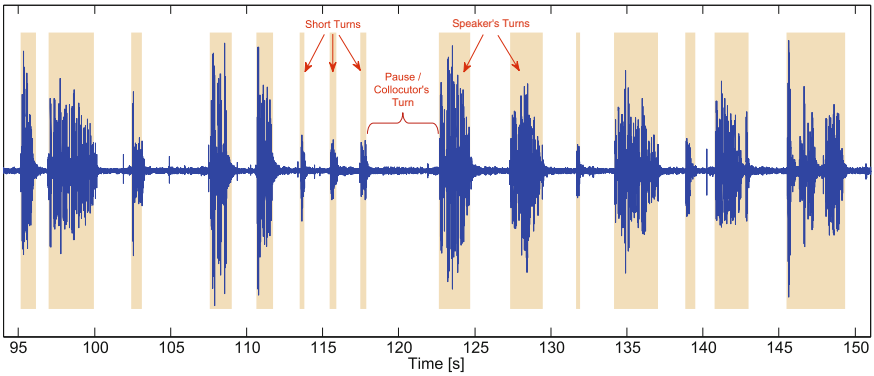

Fig. 3. Exemplary recorded smartphone audio and highlighted speaking segments

3.3 Voice Features

"openSmile" is capable of extracting more than 5000 acoustic features, but we start with a smaller feature set motivated by the findings in [5]. For each frame of the speech signal (frame length: 25 ms, step size: 10 ms) the following low-level descriptors are calculated:

- root mean square frame energy
- mel-frequency cepstral coefficients (MFCC) 1–12
- pitch frequency F_0
- harmonics-to-noise ratio (HNR)
- zero-crossing-rate (ZCR)

and to each of these, the first derivative is additionally computed. Therefore, per frame we get $16 \cdot 2 = 32$ descriptors. Next, for all frames of the speech signal the following 12 functionals are applied to the low-level descriptors:

- mean, standard deviation (2)
- kurtosis, skewness (2)
- minimum and maximum value, relative position, range (4)
- two linear regression coefficients with their mean square error (4)

Thus, the total feature vector per voice recording is $32 \cdot 12 = 384$ attributes. This high number of features is further reduced using the filter feature selection method based on joint mutual information (JMI). The JMI criterion is reported to have the best tradeoff in terms of accuracy, stability, and flexibility with small data samples [3]. Feature selection is performed using leave-one-patient-out cross-validation. Finally, we end up with the following `VOICE` features:

1. kurtosis energy
2. mean 2^{nd} MFCC
3. mean 3^{rd} MFCC
4. mean 4^{th} delta MFCC
5. max ZCR
6. mean HNR
7. std F_0
8. range F_0

4 State Recognition

4.1 Prediction Performance

The goal of the state recognition is to determine which bipolar state a patient is experiencing by using the extracted features shown in the previous chapter. To do so, we built random forest (RF) classification models for each patient individually and applied cross-validation to assess the prediction accuracy. Other classifiers were tested as well (support vector machine, neuronal network), but they achieved worse performance. In addition, RF has the built-in property to assess the importance of the variables. For each patient we applied the 3-fold cross-validation method to split randomly into training and testing sets. We chose 3 due to the small number of data samples and the unbalanced class distribution. The procedure is repeated 100 times and the mean performance values are calculated. The analysis is first carried out using only `STAT`, `SPEAK`, and `VOICE` features, and then using all features with the concatenated feature sets. Figure 4 depicts the F_1 score ($F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$) of the RF classifier using different features sets. Table 2 shows the corresponding numbers with the average values for each feature set.

The F_1 scores range from 67 % to 87 %. There is not a clear pattern that shows the best feature set, but rather, across all patients they perform similar on average (with 77 %, 78 % and 79 %). Except for patient p0902, fusing the feature sets resulted in better results with an average increase of 3 % (from 79 % to 82 %) over the best individual performances. Since the class distribution differs from patient to patient, it is important to assess the improvements above the individual baselines as well. The performance improvement ranges from low, at 9 %, for patient p102 to 19 % for patient p0902, and on average the improvement is 14 % above baseline.

The performances here are comparable with the results of the related work in [9] reporting an average accuracy of 81 % using GPS data. This performance however decreases by 5 % when GPS is fused with accelerometer features.

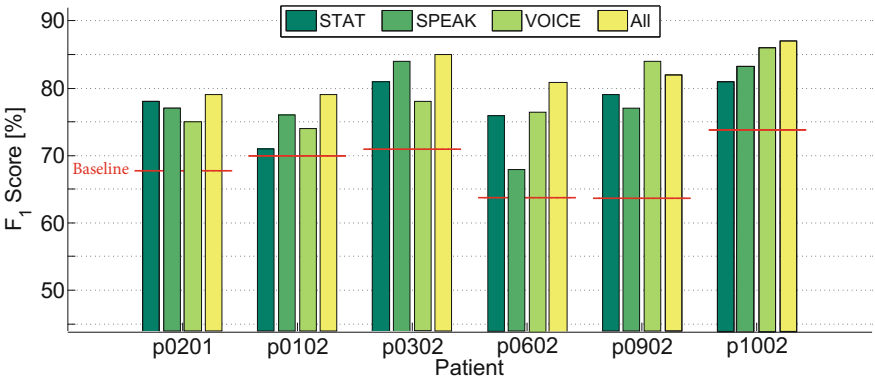

Fig. 4. Subject dependent RF performance (F_1 score) of the bipolar state classification using `STAT`, `SPEAK`, `VOICE` features and all features.

Table 2. Subject dependent RF performance (F_1 score) of the bipolar state classification and average values for each feature set.

Features	p201	p102	p302	p602	p902	p1002	Avg
`STAT`	78 %	71 %	81 %	76 %	79 %	81 %	77%
`SPEAK`	77 %	76 %	84 %	67 %	78 %	84 %	78%
`VOICE`	75 %	74 %	78 %	77 %	84 %	86 %	79%
All	**79** %	**79** %	**85** %	**81** %	**82** %	**87** %	**82**%
>Baseline	11 %	9 %	14 %	17 %	19 %	13 %	14%

4.2 Feature Significance

Beside the classification performance, we are also interested to find out which parameters are the most relevant. This side knowledge helps to better understand the behavior of a bipolar patient, and also tells us which other, non-relevant parameters can be neglected, potentially resulting in a reduced computational effort on the smartphone.

The RF classifier is able to assess the importance of the variables during the training process. In each cross-validation step (in the case of all feature sets consolidated), the importance of the features is extracted and the mean value over all steps is calculated. This value is used to sort the features in descending order. Table 3 shows the top-five features for each patient separately. The last column shows the overall top-five features, which is the weighted mean of the individual patients with the weights corresponding to the position of the features in the individual rank list.

For each patient the top-five feature ranking list varies but there are always at least two feature categories involved. This variation indicates that the patients'

behaviors are not very similar to each other, which justifies the development of person-dependent classification models. The most important average features are the average speaking length ($\texttt{SPEAK}_1$), the mean HNR value ($\texttt{VOICE}_6$), the number of short turns/utterances ($\texttt{SPEAK}_5$), the standard deviation of the pitch F_0 ($\texttt{VOICE}_7$), and the maximum daily phone call length ($\texttt{STAT}_6$).

The top features resulted from the analysis are along with important variable reported in related areas such as discriminating leader behaviour (short turns) [6], detecting stress in real-life environments using smartphones (std pitch) [10], or classifying Parkinson disease from speech (mean HNR) [15].

Table 3. Patient-wise and overall top-five important features

Rank	p201	p102	p302	p602	p902	p1002	Avg
1	$\texttt{SPEAK}_7$	$\texttt{VOICE}_3$	$\texttt{SPEAK}_2$	$\texttt{STAT}_2$	$\texttt{VOICE}_5$	$\texttt{VOICE}_7$	$\texttt{SPEAK}_1$
2	$\texttt{VOICE}_8$	$\texttt{SPEAK}_1$	$\texttt{SPEAK}_5$	$\texttt{STAT}_6$	$\texttt{VOICE}_6$	$\texttt{SPEAK}_3$	$\texttt{VOICE}_6$
3	$\texttt{STAT}_7$	$\texttt{SPEAK}_2$	$\texttt{SPEAK}_1$	$\texttt{VOICE}_7$	$\texttt{STAT}_3$	$\texttt{VOICE}_6$	$\texttt{SPEAK}_5$
4	$\texttt{STAT}_4$	$\texttt{STAT}_3$	$\texttt{STAT}_2$	$\texttt{VOICE}_8$	$\texttt{STAT}_4$	$\texttt{SPEAK}_4$	$\texttt{VOICE}_7$
5	$\texttt{VOICE}_7$	$\texttt{SPEAK}_8$	$\texttt{STAT}_6$	$\texttt{VOICE}_6$	$\texttt{SPEAK}_2$	$\texttt{VOICE}_5$	$\texttt{STAT}_6$

5 Limitations

Data Collection. From 12 patients in total we could use data from only half of. Due to the small data sample the conclusions made in this work should be treated with caution. To improve our evaluation, more subjects as well as longer trial duration are necessary.

Phone Call Features. Some features that are used in other domains could be interesting, such as the ratio of incoming/outgoing calls and the number of unique numbers, or the number of interruptions (successful and failed) during a phone call conversation.

State Recognition. In the analysis of feature significance we have shown a ranking list without giving an absolute importance weight to particular features. The assessment of the coefficients when logistic regression is used could be considered as well.

6 Conclusion and Future Work

In this work we have shown the applicability of daily phone calls to assessing the episodes of bipolar patients in a real-life environment. In order to do so,

we extracted three different types of features, namely phone call statistics, social signals derived from the phone call conversation and acoustic emotional properties of the voice.

We used the random forest classifier to train and test person-dependent models. Statistical, speaking and voice features showed on average across all patients similar individual performance in terms of state recognition. By fusing all features together, we were able to predict the bipolar states with an average F_1 score of 82 %.

Moreover, we assessed the feature importance for each person individually and we have seen that the patients behave differently from each other. Yet we identified the speaking length and phone call length, the HNR value, the number of short turns/utterances and the pitch F_0 to be the most important variables on average over all subjects.

Recognizing the current state of the bipolar patients might be difficult. However, in most cases psychiatrists are primarily interested in knowing when a person's state changes, regardless of which state the patient was in before and in what direction she/he is moving. State change triggers an alarm to the doctor, indicating that it is an important time to consult with their patient.

During the trials in Tirol we collected all data on Android smartphones. The previous work in [9] especially shows the usage of location and acceleration features for tracking bipolar states. Incorporating voice analysis could result in a complete smartphone solution for daily-life diagnosis of depressive and manic episodes in bipolar patients.

References

1. Arnrich, B., Mayora, O., Bardram, J., Tröster, G.: Pervasive healthcare - paving the way for a pervasive, user-centered and preventive healthcare model. Methods Inf. Med. **49**, 67–73 (2010)
2. Begley, C.E., Annegers, J.F., Swann, A.C., Lewis, C., Coan, S., Schnapp, W.B., Bryant-Comstock, L.: The lifetime cost of bipolar disorder in the US. Pharmacoeconomics **19**(5), 483–495 (2001)
3. Brown, G., Pocock, A., Zhao, M.-J., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J. Mach. Learn. Res. **13**, 27–66 (2012)
4. Ericsson AB. Interim Ericsson Mobility Report, February 2014. http://www.ericsson.com/ericsson-mobility-report/
5. Eyben, F., Wllmer, M., Schuller, B.: openSMILE - the Munich versatile and fast open-source audio feature extractor. In: Proceedings of ACM Multimedia (2010)
6. Feese, S., Muaremi, A., Arnrich, B., Tröster, G., Meyer, B., Jonas, K: Discriminating individually considerate and authoritarian leaders by speech activity cues. In: Workshop on Social Behavioral Analysis and Behavioral Change (SBABC) (2011)
7. Frost, M., Marcu, G., Hansen, R., Szaántó, K., Bardram, J.E.: The MONARCA self-assessment system: persuasive personal monitoring for bipolar patients. In: Proceedings of Pervasive Computing Technologies for Healthcare (PervasiveHealth) (2011)

8. Grunerbl, A., Oleksy, P., Bahle, G., Haring, C., Weppner, J., Lukowicz, P.: Towards smart phone based monitoring of bipolar disorder. In: mHealthSys (2012)
9. Grunerbl, A., Osmani, V., Bahle, G., Carrasco, J.C., Oehler, S., Mayora, O., Haring, C., Lukowicz, P.: Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In: Augmented Human (AH) (2014)
10. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T., Campbell, A.T., Perez, D.G., Choudhury, T.: StressSense: detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of ACM UbiComp (2012)
11. Merikangas, K.R., Jin, R., He, J.-P., Kessler, R.C., Lee, S., Sampson, N.A., Viana, M.C., Andrade, L.H., Hu, C., Karam, E.G., et al.: Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch. Gen. Psychiatry **68**(3), 241–251 (2011)
12. Moore, P., Little, M., McSharry, P., Geddes, J., Goodwin, G.: Forecasting depression in bipolar disorder. IEEE Trans. Biomed. Eng. **59**(10), 2801–2807 (2012)
13. Newman, S., Mather, V.G.: Analysis of spoken language of patients with affective disorders. Am. J. Psychiatry **94**(4), 913–942 (1938)
14. Osmani, V., Maxhuni, A., Grunerbl, A., Lukowicz, P., Mayora, O., Haring, C.: Monitoring activity of patients with bipolar disorder using smart phones. In: Proceedings of ACM Advances in Mobile Computing and Multimedia (2013)
15. Tsanas, A., Little, M., McSharry, P., Spielman, J., Ramig, L.: Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans. Biomed. Eng. **59**(5), 1264–1271 (2012)
16. Vanello, N., Guidi, A., Gentili, C., Werner, S., Bertschy, G., Valenza, G., Lanata, A., Scilingo, E.: Speech analysis for mood state characterization in bipolar patients. In: IEEE Engineering in Medicine and Biology Society (EMBC) (2012)
17. Xu, C., Li, S., Liu, G., Zhang, Y., Miluzzo, E., Chen, Y.-F., Li, J., Firner, B.: Crowd++: Unsupervised speaker count with smartphones. In: UbiComp (2013)
18. Young, R., Biggs, J., Ziegler, V., Meyer, D.: A rating scale for mania: reliability, validity and sensitivity. Br. J. Psychiatry **133**(5), 429–435 (1978)