

# Acoustic and Prosodic Information for Home Monitoring of Bipolar Disorder

Mireia Farrús<sup>1\*</sup>, Joan Codina-Filbà<sup>1</sup>, Joan Escudero<sup>2</sup>

<sup>1</sup> Universitat Pompeu Fabra, <sup>2</sup> Pulso Ediciones SL

[mireia.farrus@upf.edu](mailto:mireia.farrus@upf.edu), [joan.codina@upf.edu](mailto:joan.codina@upf.edu), [j.escudero@pulso.com](mailto:j.escudero@pulso.com)

## Abstract

Epidemiological studies suggest that bipolar disorder has a prevalence of about 1% in European countries, becoming one of the most disabling illnesses in working age adults, and often long-term and persistent with complex management and treatment. Therefore, the capacity of home monitoring for patients with this disorder is crucial for their quality of life. The current paper introduces the use of speech-based information as an easy-to-record, ubiquitous and non-intrusive health sensor suitable for home monitoring, and its application in the framework on the NYMPHA-MD project. Some preliminary results also show the potential of acoustic and prosodic features to detect and classify bipolar disorder, by predicting the values of the Hamilton Depression Rating Scale (HDRS) and the Young Mania Rating Scale (YMRS) from speech.

**Keywords:** bipolar disorder, home monitoring app, prosody, speech, voice

## 1. Introduction

Bipolar disorders are a common and complex form of mental disorder, ranking as one of the most disabling illnesses with a prevalence of about 1% in European countries, and up to 2.4% worldwide. Although this seems to be a small percentage, it becomes one of the most complex mental conditions. Patients with bipolar disorder experience episodes of abrupt mood changes, alternating mania (euphoria) and depression phases. But beyond these complex mental episodes, this disorder is also dynamic, with a typically relapsing-remitting course, and it becomes often a long-term and persistent illness (Strakowski et al., 2020). Moreover, patients with bipolar disorder have shown to be at high risk of premature death due to comorbid cardio-vascular diseases (Leboyer et al., 2012). In fact, several studies have demonstrated that bipolar disorder has become the sixth leading cause of disability worldwide, with a rate of death by suicide up to 15 % among the most severe cases (Marvel and Paradiso, 2004; Goodwin and Kay, 2017; Strakowski, 2014; Strakowski et al., 2020).

Since bipolar disorder alternates depression and manic phases, it is generally assessed by the research community by means of several standard clinical scales, which account for the severity of both depression and mania. The most relevant ones include the Hamilton Depression Rating Scale (HDRS) (Hamilton, 1986), and the Young Mania Rating Scale (YMRS) (Young et al., 1978). These rating scales are commonly provided by clinicians and take about 20-30 minutes to complete. YMRS rates 11 items related to mania (elevated mood, increased motor activity-energy, sexual interest, sleep, etc.) and it is based on the report from patients over the previous two days, and upon clinical observations made in a clinical interview. The scale ranges from 0 to 60, where higher scores indicate more severe mania, and its final aim is to evaluate manic state at baseline and over time. Similarly, HDRS rates 21 different items related with depression (depressed mood, feelings of guilt, suicide, insomnia, etc.), although only the first 17 compute to the final score, to indicate a degree of depression at baseline and over time. The final score ranges from 0 to 50, where higher scores indicate more severe depression.

---

\* Mireia Farrús is now with the Universitat de Barcelona

Automatic home monitoring of mood instability can allow for early intervention on prodromal symptoms and potentially influence the course of illness. Over the last years, several electronic self-monitoring platforms for regular computers and smartphones have been developed (Torous and Pawell, 2015; Gravenhorst, 2015). However, these systems usually do not have the capacity to collect objective data on patient behaviour, and most of them do not include a feedback loop between the patients and the mental healthcare providers. Recent projects such as PSYCHE [www.psyche-project.org] and MONARCA [www.monarca-project.eu] (Mayora et al., 2013) use information and communication technologies for the treatment of bipolar disorder. However, they are either intrusive for the patients due to the number of sensors needed, or they do not explore the role of the caregiver. Instead, the NYMPHA-MD project (Faurholt-Jepsen 2017) defined a framework for continuous patient monitoring, to identify early warning signs of deviations in mood and attitudes suggesting the onset of a depressive or maniac episode, which, in turn, will allow for early intervention (Codina-Filbà et al., 2021). PULSO Ediciones S.L. won the NYMPHA-MD Pre-Commercial Procurement bid with the *MoodRecord* project, where the Universitat Pompeu Fabra (Barcelona) developed the speech module analysis.

Furthermore, in recent literature, speech has been shown to be a potential indicator for bipolar disorder detection in a few works (Muaremi et al., 2014; Maxhuni et al., 2016), apart from being a ubiquitous and non-intrusive identifier. In general, systems relying on the speech signal to detect or assess mental disorders can be classified into those using acoustic-dependent features, and those using context-dependent features (Chien et al., 2019). Systems using context-dependent features require the word transcriptions to infer linguistic features (see, for instance, the use of semantic information (Mota et al., 2012) or lexical features (Voleti et al., 2019) to identify psychosis, schizophrenia and bipolar disorder). Contrarily, acoustic-dependent systems rely mainly on extracting speech-based features regardless of the linguistic content, such as spectral characteristics, voice quality features, or speech prosody. Although the context-dependent linguistic features can be very informative achieving generally better performance, they are highly dependent on language and the corresponding transcriptions from speech. The aim of the presented work is to explore the usability of speech in its acoustic form as a language-independent system for home monitoring for patients with bipolar disorder, by implementing a machine learning classifier capable of predicting the values of HDRS and YMRS scales from speech, thus providing a user-friendly tool and helpful information to patients and clinicians. Moreover, we show its integration and implementation in a smart daily-life system, specifically in the framework of the NYMPHA-MD project.

The structure of the current paper unfolds as follows. Section 2 briefly overviews the use of speech as a ubiquitous and non-intrusive health sensor. Section 3 presents some preliminary experiments on the prediction of mania and depression scales through speech-based information. Section 4 describes the MoodRecord mobile application created for patient continuous supervision in the framework of NYMPHA-MD; and finally, Sections 5 and 6 sketch the discussion and conclusions of the work, respectively.

## **2. Speech as a Ubiquitous and Non-Invasive Health Sensor**

Among all the biometric identifiers, voice has special characteristics that make it an exceptional health indicator. Moreover, speech is a non-invasive signal, ubiquitous, and easy to record, with non-expensive equipment required (Reynolds et al., 2003), which makes it especially suitable for home monitoring applications. Despite its high variability between speakers, speech is highly dependent on their physical and emotional conditions (Maltoni et al., 2003; Bolle et al., 2004), making it suitable to detect changes on these conditions.

Speech can provide two different types of information: (i) the content of the message, composed by the words and their meanings as a result of a cognitive process, and (ii) the acoustic information

extracted from the voice sound, produced by the coordinated physical activity of several organs. Variations of specific acoustic features (F0, intensity and duration) over time comprise what is known by *prosodic information*, conveyed through the intonation, stress, and rhythm elements, respectively, which can reflect the emotional aspects of the individual. In this project we focus on the second type, whose features can be extracted in a language independent manner, thus allowing a broad application to different countries with many different languages (as it was the case on the NYMPHA-MD project). In the following subsections, we introduce the potential use of speech for medical diagnosis, specifically the use of prosodic information for the analysis of the mood status and bipolar disorder.

### *2.1. Acoustic and prosodic information*

Voice has been largely used to detect several physical and mental pathologies, being acoustic parameters related to voice quality (e.g. jitter, shimmer, and harmonics-to-noise ratio) some of the most used for these purposes (Kreiman and Gerratt, 2005). By way of example, jitter and shimmer have been recently used for the detection of Parkinson's disease (Benba et al., 2014; Meghraoui et al., 2016; Sakar et al., 2017; Ali et al., 2019), Alzheimer (Mirzaei et al., 2017; Farrús and Codina-Filbà, 2020), post-traumatic stress (Banerjee et al., 2017), multiple sclerosis and dysarthria (Vizza et al., 2017; Bhat et al., 2016), thyroid patients (Gour and Udayashankara, 2015), and coronary heart disease (Pareek, 2018), among many others.

Speech prosody consists of the following elements: intonation –perceived by listeners as a variation in time of the fundamental frequency–, stress –as variation of loudness– and rhythm – as the variation of sound duration– (Adami, 2007). Prosody is crucial in oral communication (Nooteboom, 1997; Wennerstrom, 2001) and in expressing different emotional states. Furthermore, prosody has been shown to be more robust to channel noise than other spectral-based speech features (Atal, 1972), at it can be extracted in its acoustic form without the need of the corresponding text transcription.

### *2.2. Acoustic and prosodic parameters for bipolar disorder detection*

Several works in the literature have reported the usefulness of acoustic features based only on voice characteristics (from now on, *acoustic* features) to detect emotional states such as depression and mania. In (Shinohara et al., 2016), for instance, several voice quality features such as pitch rate, jitter, shimmer and harmonic-to-noise ratio indices were shown to be indicators for patients with some kind of disorders in contrast to healthy people, measured by means of a voice disability index over spontaneous speech recordings in contrast to healthy people. In (Vicsi et al., 2012), jitter, shimmer and first and second formant frequencies differed significantly in depressed speech. In (Pan et al., 2018), fundamental frequency, formants and cepstral coefficients were explored for bipolar detection. In a similar way, Shimizu et al. (2005) analysed the chaotic behaviour of vocal sounds in patients with depression, and Zhou et al. (2001) proposed a new feature value based on the nonlinear Teager energy operator to classify speech under stressed conditions. Other voice quality characteristics were further studied in (Scherer et al., 2013) and (Hargreaves et al., 1965) to detect depression and post-traumatic stress disorder, and it has also been shown that pressure of speech is a powerful indicator for mania states (Carlson et al., 1973).

The use of prosodic features also in its acoustic form (from now on *prosodic* features), has been shown to be highly relevant to identify emotions (Luengo et al., 2005; Mary, 2019), which leads to think of its usefulness in the detection of bipolar disorder –although the literature in this field is less exhaustive. See, for instance, the use of intonation (pitch contour) in (Guidi et al., 2015), rhythm features in (Gideon et al., 2016).

### 3. Bipolar Mood Status Detection from Acoustic and Prosodic Information

The current section presents some preliminary experiments on the use of speech-based features for the detection of bipolar disorders. Since the data collected were not exhaustive, the aim of these experiments is to show the potential of speech features to classify individuals' speech into depression and mania scales. The following subsections describe the gathering of data and feature extraction in the MoodRecord application, the setup for bipolar status detection, and the corresponding preliminary results.

#### 3.1. Data recording and feature extraction

The original recordings were acquired by two of the EU procurers working in the project, namely *Consorci Sanitari Parc Taulí* (CSPT, Barcelona), and *Provincia Autonoma di Trento* (PAT, Trento). The recordings were carried out by bipolar patients gathered from the two institutions, using the MoodRecord application, and further processed with Praat, a free computer software package for speech analysis (Boersma and Weenink, 2017). The recordings had an average duration of 26.0 seconds, and an average effective duration (excluding silences) of 18.5 seconds.

We developed a Praat-based module to extract acoustic and prosodic features from the sound files. These features were then used to train machine learning models to detect mania, depression and normal states by means of YMRS and HDRS scales. The selection of the acoustic features was mainly based on existing literature. Thus, several works, as seen in previous sections, have reported the use of fundamental frequency (F0), formants, and voice quality features. Other few works have also reported the use of prosodic features based on intonation and rhythm. We used a similar set of acoustic features, which lead to the following nine features:

- fundamental frequency (1 feature): mean value of F0
- formants (2 features): first formant frequency (F1), and second formant frequency (F1)
- voice quality (6 features): relative value of jitter, absolute value of jitter, relative value of shimmer, and relative value of shimmer (Farrús and Hernando, 2009), noise-to-harmonics ratio (NHR), and harmonics-to-noise ratio (HNR)

Moreover, we also extracted nine prosodic features, based on the following three prosody elements:

- intonation (4 features): maximum value of F0, minimum value of F0, range of F0, slope of F0
- stress (1 feature): mean value of intensity
- rhythm (4 features): ratio of pauses, speech rate, articulation rate, and average syllable duration (De Jong and Wempe, 2009)

In total, we extracted 18 features from each of the patients' speech. For the extraction of F0 and its related features, we used the auto-correlation method in Praat with an interval of 10 ms and a Hanning window of length 40 ms. Although rhythm features could have been accurately extracted from the corresponding transcripts, they were extracted by adapting the Praat script found in (De Jong and Wempe, 2009), to keep the system language independent. The mean F0 and intensity values were used to normalize the F0- and intensity-based features, respectively, to avoid speaker dependence. Thus, F0-based features were computed as distance in semitones with respect to the mean value of the individual. Voice quality parameters jitter and shimmer were also normalized by means of F0 and intensity, respectively. After normalization, F0 and intensity were left out and the remaining 16 features (eight voice quality and formant features plus eight prosodic features) were used for the detection experiments.

### 3.2. Bipolar status detection

The detection of mood status in MoodRecord application was performed by means of regression algorithms using the extracted speech features. To this end, doctors annotated the different speech recordings according to YMRS and HDRS scales so that the system can learn from such scores to point out the occurrence of mania or depressive episodes. Specifically, the detection of depression and mania states was based on the range of scales and mania/depression assessments (Martino et al., 2017) indicated in Table 1, considering also the case of “no mania” and “no depression” (euthymic state). The scales were assigned to those audios yielding within three days before and after the date of the doctor’s assessment and thus the corresponding scale assignment.

Depression	HDRS value	Mania	YMRS value
no	$< 4$	no	$< 4$
subclinical	$5 \leq \text{HDRS} < 10$	subclinical	$5 \leq \text{YMRS} < 9$
mild	$10 \leq \text{HDRS} < 16$	mild	$9 \leq \text{YMRS} < 16$
moderate	$16 \leq \text{HDRS} < 25$	moderate	$16 \leq \text{YMRS} < 25$

Table 1. Ranges of HDRS and YMRS scales in different mania and depression states.

### 3.3. Preliminary experiments

For an initial testing of the system, some preliminary regression experiments were performed within the framework of the project. An initial small database consisting of 65 recordings from 19 different real users and patients aged 23-69 years old was collected in the framework of the NYMPHA-MD project (see more details on patient recruitment in (NYMPHA, 2018)). After a careful check on the quality of the audios, only 49 of them –corresponding to 13 different users– were found to be recorded with enough quality. The remaining ones were too short –less than two seconds–, empty, or containing only noise. Moreover, 15 recordings were shorter than 5 seconds, which could be used to extract acoustic features but were not suitable for computing reliable prosodic features over time, so they were not annotated with their corresponding HDRS and YMRS scales by the clinicians. For other recordings, clinicians were not available at the time of recordings. Since assessments had to be done *in situ* and within a short time range with respect to the recording time to be reliable, only 30 out of the valid audios could be annotated. Furthermore, the variability of the scales was rather low: a great majority of them were associated to “no depression” and “no mania” scales. Table 2 summarises these statistics.

	# recordings	diff. users
<b>Total recordings</b>	65	19
<b>Valid recordings</b>	49	13
<b>Valid recordings with scales</b>	30	11
<b>States</b>		
no depression /no mania	24	7
subclinical depression / no mania	2	1
mild depression / no mania	0	0
moderate depression /no mania	1	1
no depression /subclinical mania	3	2
no depression /mild mania	0	0
<b>no depression / moderate mania</b>	0	0

Table 2. Statistics of the speech recordings within the NYMPHA-MD project.

The regression experiments were performed on Weka (Frank et al. 2016), using a leave-one-out (LOO) cross-validation method to ameliorate the lack of data. To predict the HDRS and YMRS

values, three different regression algorithms were tested: linear regression, random forest, and support vector regression with a radial kernel. Tables 3 and 4 show the results obtained in the prediction of HDRS and YMRS scales, respectively, in terms of root mean square error (RMSE) and using the following sets of features: (1) voice quality and formants, (2) prosody, and (3) all features. Bold numbers represent the best-performing set of features for each classifier. Moreover, Table 5 compares the best results obtained in the LOO for both HDRS and YMRS scales with 10- and 5-fold cross-validations. The results clearly show that the larger the number of folds, the better the accuracy achieved in the regression experiments.

	Linear Regression	Random Forest	Support Vector Regression
Voice quality + formants	<b>4.275</b>	<b>4.398</b>	<b>3.945</b>
Prosody	5.052	4.545	4.020
All	5.063	4.604	4.049

Table 3. Root mean square errors obtained in the prediction of HDRS values (LOO cross-validation).

	Linear Regression	Random Forest	Support Vector Regression
Voice quality + formants	2.226	2.067	2.244
Prosody	2.497	2.075	2.189
All	<b>1.985</b>	<b>2.021</b>	<b>2.156</b>

Table 4. Root mean square errors obtained in the prediction of YMRS values (LOO cross-validation).

	cross-validation	Linear Regression	Random Forest	Support Vector Regression
HDRS Voice quality + formants	LOO	<b>4.275</b>	<b>4.398</b>	<b>3.945</b>
	10-fold	5.620	4.617	4.269
	5-fold	4.441	4.935	4.115
YMRS All features	LOO	<b>1.985</b>	<b>2.021</b>	<b>2.124</b>
	10-fold	2.329	2.045	2.134
	5-fold	2.671	2.172	2.139

Table 5. RMSE values obtained using 5-, 10-fold and LOO cross-validation.

On the one hand, Table 3 shows that the best HDRS prediction is obtained by using only the voice quality plus formants set of features. In addition, among all the regression methods tested, support vector regression algorithm outperformed the other two classifiers. On the other hand, Table 4 shows that the best YMRS prediction is obtained with the whole range of speech features. Unlike the HDRS prediction, prosody here provides useful information for the YMRS detection, which could be explained by the fact that manic speech is conveyed in a much higher degree by prosody, with higher variations in intonation and rhythm than depressed speech. In Figures 1 and 2 we

have plotted the actual and predicted values for both HDRS and YMRS, respectively, for the 30 speech samples corresponding to the lowest RMSE obtained for each scale. The graphics show that the most extreme values are more difficult to be predicted due to the lack of representative training data. These results should be interpreted in caution due to the minimal number of valid recordings obtained. Moreover, most of the valid recordings were very close to the euthymic states, which limits the variability of the data used in the experiments. Note that, in the final application, since negative values make no sense in a real setting, negative predicted values are turned into 0 values. In the same way, large values are cut to the maximum HDRS and YMRS values (50 and 60, respectively).

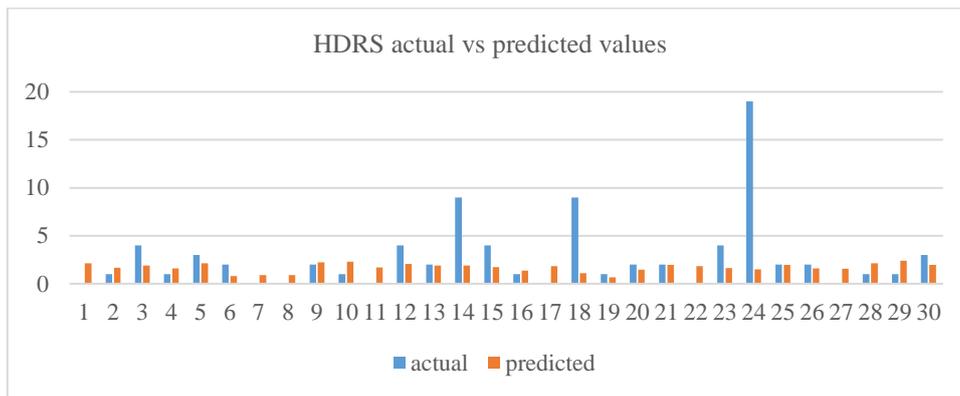


Figure 1. Actual HDRS values compared to the HDRS predicted values using voice quality and formant features through a support regression algorithm.

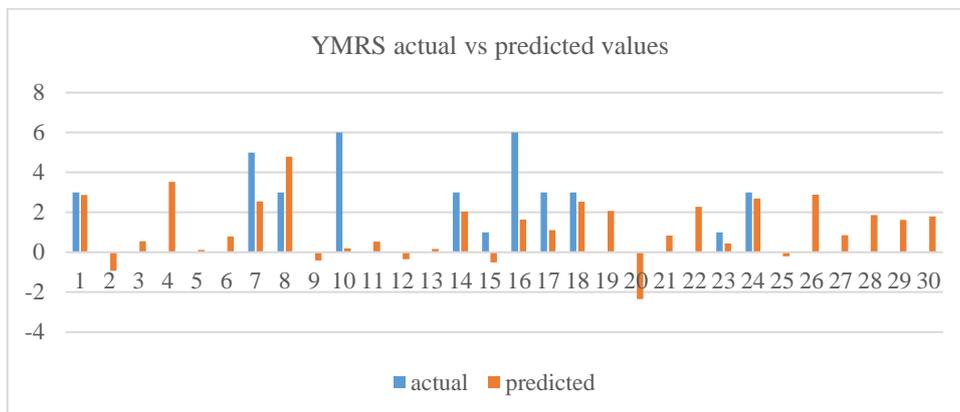


Figure 2. Actual YMRS values compared to the YMRS predicted values using all speech features through a linear regression algorithm.

#### 4. Patient Continuous Supervision through Mobile App

The application developed within the NYMPHA-MD aimed to provide a new way to manage patients diagnosed with bipolar disorder through the *MoodRecord* system that allows the estimation of the mood of the patient and the patient monitoring.

##### 4.1. MoodRecord System

The MoodRecord System was designed to be used for patients (app functionalities) and healthcare professionals and caregivers (website functionalities, [www.moodrecord.com](http://www.moodrecord.com)). The patient registers a set of parameters related to their mood using the app. The website provides clinicians with all user's data registered by the app to manage and track their patients' disorders (Figure 3).

The flow starts when clinician registers new patients from the web interface and sends them their credentials. Patients are then able to access the application on Android or IOS smartphones and start registering the data related to their mood or health state following the weekly guideline defined by their case manager (see in Figure 3 the diagram of the MoodRecord system architecture).

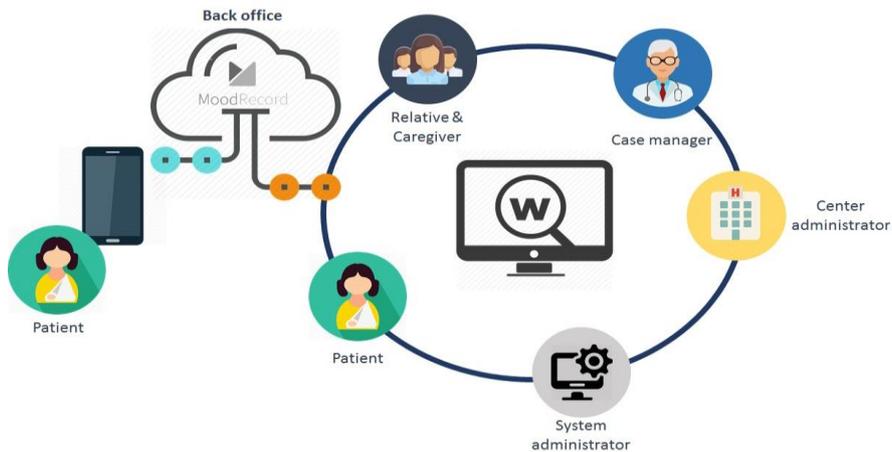
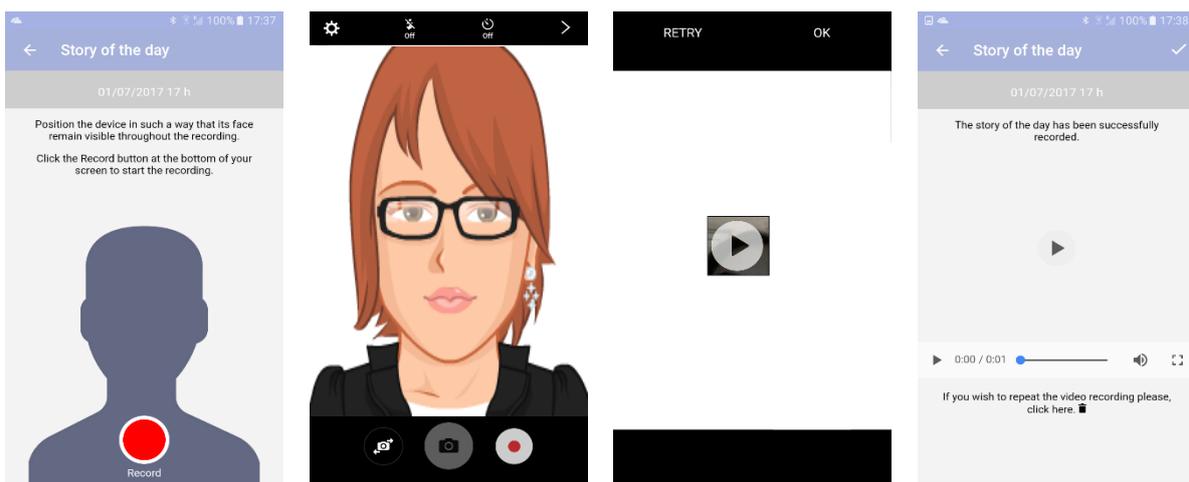


Figure 3. Diagram of the MoodRecord system architecture.

The speech recordings in MoodRecord are performed through a module called *Story of the day*. In this module, the patients are asked to record a video explaining their day, and it includes two functionalities: face recognition, and speech pattern detection. The system is first calibrated to remove the microphone noise. After that, the *Story of the day* module is ready for the video/audio recording. Figure 4 shows the different steps followed in this process.



**Step 1:** Read the instructions on the screen and click on the Record button.

**Step 2:** Start recording the *Story of the day* explaining recording and click on “OK” what you have done and how you have felt.

**Step 3:** Once finished, stop the video, click on the “tick” button to save it.

**Step 4:** Play the recorded video, click on the “bin” button to send it or on the “bin” icon to repeat the process from first step.

Figure 4. *Story of the day* registration.

The recorded video is then sent for speech analysis. Audio is separated from image and speech features are extracted for their analysis and further mania and depression scales prediction, based on the manually annotated speeches previously used to train the system. The system is initially built using the normalised speech features to develop a generic classifier. As training data is being collected, the system can be focused on the development of a personalised user model.

#### *4.2. Medical supervision*

The clinician sets up initial and follow-up visits, in which the patient status is revised and assessed again. Between such doctor visits, patients use the MoodRecord app to record a video through the *Story of the day* module so that they can be home monitored. Apart from audio and facial recognition algorithms, the system includes other patient parameters such as sleep quality, personal questionnaires, etc., which are checked online by the clinician. The system includes alarms that are raised when one of the indicators achieve a critical value.

### **5. Discussion**

The preliminary experiments presented in the previous sections have shown the feasibility of voice features to detect bipolar status, since audio features can reflect both manic and depressive states. With respect to other existing works, the results are comparable when proving the usefulness of speech in bipolar status detection. Muaremi et al. (2014), for instance, show that voice features are objective markers of emotional states in bipolar disorder –improved when combined with other data extracted from smartphones–, and that they become more effective in the detection of mania than the detection of depression. Maxhuni et al. (2016), on the other hand, test both prosodic and spectral speech features and find that both types have similar accuracy when tested together or in isolation. Our analysis goes beyond these works by testing the specific contribution of both prosodic and other acoustic features, and finds that, while prosody provides useful information for the YMRS detection (manic state), HDRS detection (depressive state) relies more on the non-prosodic acoustic features.

Our experiments could be improved by collecting more data to produce better machine learning models. Moreover, the short time used to test the application with patients implies that patients have been stable during the recorded period, and that they have not had any significant changes in their mood states –often, the mood status can change when turning into different seasons, so a collection period over six months would be needed for a better performance in the prediction algorithm–. The low data variability is a handicap for training the models. To overcome it, data collected in the future will automatically be used to improve both the individual and the generic models.

### **6. Conclusion**

In this work, we have presented the use of acoustic and prosodic information as a health indicator; concretely, as an identifying factor of bipolar disorders. Speech is ubiquitous and easy to record, which makes it a suitable identifier for home monitoring systems. Some preliminary experiments on the use of acoustic and prosodic features in the framework of the NYMPHA-MD project have shown promising capabilities to detect different mood states from speech, overcoming those systems based only on voice characteristics, without an extra overload for the patient. Moreover, the MoodRecord application presented in the current work becomes a practical tool for a further medical supervision of the patient, leveraging the need of regular physical visits between patients and clinicians.

Although preliminary, the results show that, within the range of the available data, our speech-based algorithm is a potential tool for predicting different mood status in patients with bipolar disorders. Therefore, the use of speech and more linguistic-prosodic information should be thus further included in home monitoring health systems.

## 7. Acknowledgements

The first author has been funded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under grant RYC-2015-17239 (AEI/FSE, UE). This work is part of the MYMPHA-MD project, which has been funded by the European Union under Grant Agreement N° 610462.

## 8. References

Adami AG (2007) Modeling prosodic differences for speaker recognition. *Speech Communication* 49(4): 277-291.

Ali H, Adnan SM, Aziz S, Ahmad W and Obaidullah M (2019) Sound classification of Parkinsonism for telediagnosis. *Technical Journal* 24(1): 90-97.

Atal BS (1972) Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America*, 52(6B): 1687–1697.

Banerjee D, Oslam K, Mei G, Xiao L, Zhang G, Xu R, Ji S and Li J (2017) A deep transfer learning approach for improved post-traumatic stress disorder diagnosis. In: *IEEE International Conference on Data Mining*, pp. 11-20.

Benba A, Jilbab A and Hammouch A (2014) Hybridization of best acoustic cues for detecting persons with Parkinson's disease. In: *World Conference on Complex Systems*, pp. 622-625.

Bhat C, Vachhani B, and Kopparapu SK (2016) Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation. In: *Proceedings of Interspeech*, pp. 228-232.

Boersma P and Weenink D (2017) Praat: doing phonetics by computer [Computer program]. Version 6.0.29, retrieved February 2017 from <http://www.praat.org/>

Bolle RM, Connell JH, Pankanti S, Ratha NK, and Senior AW (2004) *Guide to Biometrics*. Springer, New York.

Carlson GA and Goodwin FK (1973) The stages of mania: A longitudinal analysis of the manic episode. *Archives of General Psychiatry* 28(2): 221-228.

Chien Y, Hong S, Cheah W, Yao LH, Chang YL, Fu C (2019) An Automatic Assessment System for Alzheimer's Disease Based on Speech Using Feature Sequence Generator and Recurrent Neural Network. *Sci Rep* 9, 19597. <https://doi.org/10.1038/s41598-019-56020-x>

Codina-Filbà J, Escalera S, Escudero J, Antens C, Buch-Cardona P, Farrús M. Mobile eHealth platform for home monitoring of bipolar disorder. In Proceedings of the 27th International Conference on Multimedia Modeling (MMM), Prague, Czech Republic, 2021.

De Jong NH and Wempe T (2009) Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41(2): 385-390.

Farrús M and Codina-Filbà J (2020) Combining Prosodic, Voice Quality and Lexical Features to Automatically Detect Alzheimer's Disease. *Submitted to Interspeech 2020*.

Farrús M and Hernando J (2009) Using jitter and shimmer in speaker verification. *IET Signal Processing* 3(4): 247-257.

Faurholt-Jepsen M (2017) The NYMPHA-MD project: Next generation mobile platforms for health, in mental disorders. *European Psychiatry* 41: S23.

Frank E, Hall MA and Witten IH (2016) The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.

Gideon J, Provost EM, and McInnis M (2016) Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2359-2363.

Gravenhorst F, Muaremi A, Bardram J, Grünerbl A, Mayora O, Wurzer G, ... & Tröster G (2015). Mobile phones as medical devices in mental disorder treatment: an overview. *Personal and Ubiquitous Computing*, 19(2), 335-353.

Goodwin FK, Kay RJ (2017) *Manic-depressive illness: bipolar disorders and recurrent depression*. Vol. 1. Oxford University Press.

Gour GB and Udayashankara V (2015) Voice disorder analysis of thyroid patients. *International Journal of Computer Science and Information Technology* 4(5): 720-727.

Guidi A, Vanello N, Bertschy G, Gentili C, Landini L, and Scilingo EP (2015) Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients. *Biomedical Signal Processing and Control*, 17, 29-37.

Hamilton M (1986) The Hamilton Rating Scale for Depression. *Assessment of Depression*, pp. 143-152.

Hargreaves WA, Starkweather JA and Blacker KH (1965) Voice quality in depression. *Journal of Abnormal Psychology* 70(3): 218.

Kreiman J and Gerratt BR (2005) Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America* 117(4): 2201–2211.

Leboyer M, Soreca I, Scott J, et al. Can bipolar disorder be viewed as a multi-system inflammatory disease? *J Affect Disord*. 2012;141(1):1-10. doi:10.1016/j.jad.2011.12.049

Luengo I, Navas E, Hernández I, and Sánchez J. (2005). Automatic emotion recognition using prosodic parameters. In *Ninth European Conference on Speech Communication and Technology*.

Maltoni D, Maio D, Jain AK, and Prabhakar S (2003) *Handbook of Fingerprint Recognition*. Springer, New York.

Marvel CL, Paradiso S. Cognitive and neurological impairment in mood disorders. *Psychiatr Clin North Am*. 2004;27(1):19-viii. doi:10.1016/S0193-953X(03)00106-0

Martino DJ, Samamé C, Marengo E, Igoa A, Scápola M, Strejilevich SA (2017) A longitudinal mirror-image assessment of morbidity in bipolar disorder. *Eur. Psychiatry*, 40: 55-59.

Mary, L (2019) Extraction and Representation of Prosody for Speaker, Language, Emotion, and Speech Recognition. *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*. Springer, Cham, pp. 23-43.

Maxhuni A, Muñoz-Meléndez A, Osmani V, Pérez H, Mayora O, and Morales EF (2016) Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing*, 31: 50-66.

Mayora O, Arnrich B, Bardram J, Dräger C, Finke A, Frost M, ... and Haux R (2013). Personal health systems for bipolar disorder: Anecdotes, challenges and lessons learnt from MONARCA project. In: *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, pp. 424-429.

Meghraoui D, Boudraa B, Merazi-Meksen T, Boudraa M (2016) Parkinson's disease recognition by speech acoustic parameters classification. *Modelling and Implementation of Complex Systems*, pp. 165-173.

Mirzaei S, El Yacoubi M, Garcia-Salicetti S, Boudy J, Muvingi CKS, Cristancho-Lacroix V, Kerhervé H, and Rigaud-Monnet AS (2018) Two-stage features selection of voice parameters for early Alzheimer's disease prediction. In: *IRBM* 39(6): 430-435.

Mota NB, Vasconcelos NA, Lemos N, Pieretti AC, Kinouchi O, Cecchi GA, ... and Ribeiro S (2012) Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4).

Muaremi A, Gravenhorst F, Grünerbl A, Arnrich B., and Tröster G. (2014) Assessing bipolar episodes using speech cues derived from phone calls. In: *International Symposium on Pervasive Computing Paradigms for Mental Health*, pp. 103-114.

Nooteboom S (1997) *The Prosody of Speech: Melody and Rhythm. The Handbook of Phonetic Sciences*. Blackwell Publishers Ltd, Oxford.

NYMPHA (2018) NYMPHA-MD: Next Generation Mobile Platforms for HeAlth, in Mental Disorders. *Technical Specification*, retrieved from: [http://www.appalti.provincia.tn.it/binary.php/pat\\_pi\\_bandi\\_new/bandi/NYMPHA\\_MD\\_tecnical\\_specification.1441203112.pdf](http://www.appalti.provincia.tn.it/binary.php/pat_pi_bandi_new/bandi/NYMPHA_MD_tecnical_specification.1441203112.pdf)

Pareek V (2018) Detection of coronary heart disease from vocal profiling and to determine the vocal tract transfer function and glottal excitation pulse. Technical report, National Institute of Technology, Kurukshetra.

Pan Z, Gui C, Zhang J, Zhu J, and Cui D (2018) Detecting manic state of bipolar disorder based on support vector machine and gaussian mixture model using spontaneous speech. *Psychiatry Investigation*, 15(7), 695.

Reynolds DA, Andrews W, Campbell J, Navratil J, Peskin B, Adami A, Qin Jin, Klusacek D, Abramson J, Mihaescu R, Godfrey J, Jones D, and Bing Xiang. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In: *Proceedings of the ICASSP*, vol. 4, pp. 784–787, Hong Kong, China.

Sakar BE, Serbes G and Sakar CO (2017) Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PloS one* 12/8, e0182428.

Scherer S, Stratou G, Gratch J, Morency LP (2013) Investigating voice quality as a speaker-independent indicator of depression and PTSD. In: *Proceedings of Interspeech*.

Shimizu T, Furuse N, Yamazaki T, Ueta Y, Sato T, Nagata S (2005) Chaos of vowel /a/ in Japanese patients with depression: a preliminary study. *Journal of Occupational Health* 47(3): 267-269.

- Shinohara S, Nakamura M, Mitsuyoshi S, Tokuno S, Omiya Y and Hagiwara N (2017) Voice disability index using pitch rate. In: *IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*.
- Strakowski, S. M. (2014). A programmatic approach to treatment. *Bipolar Disorder*. Oxford American Psychiatry Library.
- Strakowski SM, Fleck DE, Adler CM, DelBello MP (eds) (2020). *Bipolar Disorder*. Oxford University Press.
- Torous J and Powell AC (2015) Current research and trends in the use of smartphone applications for mood disorders. *Internet Interventions*, 2(2): 169-173.
- Vicsi K, Sztahó D and Kiss G (2012) Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. In: *IEEE 3<sup>rd</sup> International Conference on Cognitive Infocommunications (CogInfoCom)*.
- Vizza P, Mirarchi D, Tradigo G, Redavide M, Bossio RB and Veltri P (2017) Vocal signal analysis in patients affected by Multiple Sclerosis. *Procedia Computer Science* 108: 1205-1214.
- Voleti R, Woolridge S, Liss JM, Milanovic M, Bowie CR, and Berisha V (2019) Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder. arXiv preprint arXiv:1904.10622.
- Wennerstrom A (2001) *The Music of Everyday Speech. Prosody and Discourse Analysis*. Oxford University Press.
- Young RC, Biggs JT, Ziegler VE, and Meyer DA (1978) A Rating Scale for Mania: Reliability, Validity and Sensitivity. *British Journal of Psychiatry*, vol. 133, pp. 429-435.
- Zhou G, Hansen JHL and Kaiser JF (2001) Nonlinear feature-based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing* 9(3): 201-216.