Discrimination of Bipolar Disorders Using Voice

Masakazu Higuchi[1(✉)], Mitsuteru Nakamura[1], Shuji Shinohara[2], Yasuhiro Omiya[3], Takeshi Takano[3], Hiroyuki Toda[4], Taku Saito[4], Aihide Yoshino[4], Shunji Mitsuyoshi[2], and Shinichi Tokuno[1]

[1] Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
{higuchi,m-nakamura,tokuno}@m.u-tokyo.ac.jp
[2] Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
{shinohara,mitsuyoshi}@bioeng.t.u-tokyo.ac.jp
[3] PST Inc., Yokohama, Japan
{omiya,takano}@medical-pst.com
[4] Department of Psychiatry, National Defense Medical College, Tokorozawa, Japan
{toda1973,tsaito,aihide}@ndmc.ac.jp

Abstract. Several methods have been developed for screening mentally impaired patients using biomarkers, but these methods are invasive and costly. Self-administered tests are also used as screening methods. They are non-invasive and relatively simple, but they cannot eliminate the influence of reporting bias. On the other hand, the authors have conducted studies on technologies for inferring the mental state of persons from their voices. Analysis using voice has the advantage of being non-invasive and easy to perform. This study proposes a vocal index that will distinguish between a healthy person and a bipolar I or II patient using a polytomous logistic regression analysis with patients with bipolar disorder as subjects. When the subjects were classified using the prediction model obtained from the analysis, the subjects were categorized into three groups with an accuracy of approximately 67%. This result suggested that the vocal index could be a new evaluation index for discriminating between subjects with and those without bipolar disorder.

Keywords: Voice · Bipolar disorders · Polytomous logistic regression analysis

1 Introduction

In terms of the screening of patients with mental illness, methods have been developed that use biomarkers [1–3] such as saliva, blood, and heart rate; however, they are invasive and costly. Noninvasive methods include self-administered psychological tests [4–6] such as the General Health Questionnaire (GHQ), Beck Depression Inventory (BDI), and Young Mania Rating Scale (YMRS), which are used commonly. Although self-administered tests are relatively easy, they cannot completely eliminate reporting bias. Reporting bias occurs when certain

P. Cipresso et al. (Eds.): MindCare 2019, LNICST 288, pp. 199–207, 2019.
https://doi.org/10.1007/978-3-030-25872-6_16

pieces of information are selectively under or over-evaluated, either consciously or subconsciously, by the respondents [7].

On the other hand, a change in mood is empirically known to be manifested in the voice, with research studies having been conducted to infer depressive or stressed states of subjects using voices [8–11]. Assessment using the voice has a number of advantages. It is noninvasive, allows for easy and remote analysis, and does not require special, specific devices. Furthermore, it also has the potential to resolve several issues with detecting psychiatric disorders, including the reporting bias in self-administered psychological testing.

Bipolar disorder is a psychiatric disorder in which the patient goes through a cycle of manic and depressive episodes, with type I involving stronger manic states and type II involving milder ones [12]. Diagnosing bipolar disorder is often difficult, even for experts. In particular, it is difficult to distinguish depression caused by bipolar disorder from unipolar depression. The earlier the age of onset, the higher the likelihood of the first few episodes to be depressive [13]. Since the diagnosis of bipolar disorder requires a manic or hypomanic episode, many patients are initially misdiagnosed as having major depression [14]. Therefore, newer technology that can distinguish between major depression and bipolar disorder at an early stage is required. We have been conducting research on a voice index that can detect bipolar and major depressive disorders in patients [15,16]. In a study with only bipolar patients, Faurholt-Jepsen et al. presented that the manic state measured using the YMRS can be discriminated with high accuracy from vocal features [17]. Maxhuni et al. reported that it is possible to classify with high confidence the course of mood episodes or relapse in bipolar patients, using motor activity information including audio, accelerometer and self-assessment data [18]. However, neither study has conducted an analysis by separating type I and II bipolar disorders.

Therefore, this study only examined patients with bipolar disorder to propose a vocal evaluation index that will differentiate between type I and II bipolar disorders using healthy subjects as a control.

2 Methods

2.1 Subjects

The voices of patients who visited the National Defense Medical College Hospital for the treatment of bipolar disorder were studied, as well as those of healthy subjects, who lived their everyday lives without mental health issues. Of the patients, 25 had type I bipolar disorder (BPI) and 39 had type II bipolar disorder (BPII). The patients were diagnosed using the Mini-International Neuropsychiatric Interview (MINI) [19]. Moreover, the patients were interviewed by a doctor using Hamilton Depression Rating Scale (HAM-D) [20], and a self-administered psychological test, YMRS, was conducted. Fourteen healthy subjects (HE) were included.

2.2 Acquisition of Voices

Vocal data were acquired by recording the voices of the subjects, reading out a fixed sentence comprised of 17 Japanese phrases, after obtaining their consent. This reading aloud was conducted twice. The vocal recordings were conducted in the hospital consultation room for both the healthy subjects, and the patients. The voice was recorded using the pin microphone ME52W (Olympus, Tokyo, Japan) attached at the breast, approximately 10 cm away from the subject's mouth. The recording device used was the Portable Recorder R-26 (Roland, Shizuoka, Japan), with the recording format being 96 kHz and 24-bit linear PCM.

2.3 Analysis of Voice Data

Selection of Features. The vocal features extracted from voices of subjects. The feature extraction was conducted using the freeware openSMILE version 2.3 [21]. The openSMILE freeware comes with scripts that automatically extract vocals from various feature sets. In this study, the feature set used in emotion recognition (the large openSMILE emotion feature set) was extracted from each voice. From a single vocal data, 6,552 vocal features were extracted. From these features, those that fit the model were selected. The procedure used was as follows:

1. Of the data, 75% were extracted from the HE, BPI, and BPII groups, respectively, using feature data of all subject's vocal data, with each used as a data set for training. The remaining 25% of the data were used as the data set for testing the model. The details for each data set are presented in Tables 1 and 2.
2. For feature f, the training data were divided into the HE_f, BPI_f, and BPII_f groups, with the combination of each pair of groups being HE_f versus BPI_f, HE_f versus BPII_f, and BPI_f versus BPII_f. The classification performance of the two groups were calculated for all combinations using the area under the curve (AUC) of the receiver-operating characteristic (ROC). In this study, a feature in which the AUC was >0.9 for combinations HE_f versus BPI_f or HE_f versus BPII_f was selected, or a feature of an AUC of >0.7 for the combination BPI_f versus BPII_f was selected.
3. A correlation analysis was conducted for the features of the training data selected in step 2. With a feature pair in which the Pearson product-moment correlation coefficient exceeded 0.8, one of the features was eliminated.

Multivariate Analysis. By using the features of the selected training data as the explanatory variable and category information as objective variables, a regularized polytomous logistic regression analysis was conducted [22]. The polytomous logistic regression analysis is a multivariate analysis method that categorizes data into three groups or more, and is an extended version of the regular logistic regression analysis that categorizes data into two groups. From the model formula obtained from the analysis, the probability of each piece of data belonging to each group is estimated, with the data categorized into the

group with the highest probability. The probability P_g in which the data x comprising of F pieces of features belonging to group $g \in \{\text{HE, BPI, BPII}\}$ is calculated using the following equation:

$$P_g = \frac{\exp\left(\alpha_g + \sum_{i=1}^{F} \beta_{ig} x_i\right)}{\sum_{j \in \{\text{HE, BPI, BPII}\}} \exp\left(\alpha_j + \sum_{i=1}^{F} \beta_{ij} x_i\right)}, \tag{1}$$

where the α_g, $\beta_{1g}, \ldots, \beta_{Fg}$ are the model coefficient in the model formula for group g. For statistical processing, the statistical analysis freeware R version 3.4.2 [23] was used.

Evaluation. The model performance was evaluated by calculating the precision (the percentage of correctly predicted subjects among the subjects predicted as a group), recall (the percentage of correctly predicted subjects among the subjects in a group), and accuracy (the percentage of correctly predicted subjects among the total subjects) from the confusion matrix for both training and test data.

Table 1. Subjects' information of the training data set.

Category		Number of subjects	Age	HAM-D score	YMRS score
HE	Male	7	42.0 ± 4.6 (n/a 2)	-	-
	Female	4	28 (n/a 3)	-	-
BPI	Male	10	56.3 ± 11.1	5.4 ± 5.8	0.6 ± 1.3
	Female	9	57.9 ± 15.0	5.9 ± 7.3	0.8 ± 1.7
BPII	Male	12	52.0 ± 14.4	6.2 ± 6.0	1.6 ± 3.3
	Female	18	51.1 ± 12.4	7.2 ± 7.4	2.1 ± 2.8

"n/a" signifies the missing value of the sample.

Table 2. Subjects' information of the test data set.

Category		Number of subjects	Age	HAM-D score	YMRS score
HE	Male	3	47.0 ± 11.3 (n/a 1)	-	-
	Female	0	-	-	-
BPI	Male	3	45.3 ± 13.3	3.3 ± 3.5	1.3 ± 2.3
	Female	3	56.0 ± 14.9	1.0 ± 1.0	2.3 ± 4.0
BPII	Male	2	36.5 ± 9.2	8.0 ± 4.2	2.5 ± 0.7
	Female	7	55.3 ± 11.0	5.0 ± 4.2	2.7 ± 4.3

"n/a" signifies the missing value of the sample.

3 Results

3.1 Features of the Model

As a result of step 2 of feature selection, 402 pieces of data were collected from 6,552 features. As a result of step 3 of the feature selection, 55 pieces of data were collected from 402 feature volumes. As a result of polytomous logistic regression analysis, a prediction model comprised of 28 features in total was obtained. For the individual model formulas, they were probability prediction formulas comprised of 6 features for the HE group, 13 features for the BPI group, and 11 features for the BPII group. Moreover, the model included features related to low to middle frequencies, Mel-frequency, signal power, zero crossing rate of time domain, and voiced sound.

3.2 Performance of the Model

Performance for Training Data. The confusion matrix in Table 3 was obtained as a result of conducting predictions on the training data using the prediction model. Over 90% recall was found in all the groups, with >90% precision found in all prediction groups, resulting in >95% accuracy found for the overall data.

Table 3. The confusion matrix of the training data according to prediction model.

		Prediction				
		HE	BPI	BPII	Recall	
Actual	HE	10	0	1	90.9%	
	BPI	0	18	1	94.7%	
	BPII	0	0	30	100%	
	Precision	100%	100%	93.8%	Accuracy	96.7%

The distribution of group discrimination probability of the subjects who belonging to each group is presented in Fig. 1. In all groups, the probability that the subjects were classified into the same group as the one they originally belonged tended to be higher than the probability that the data were classified into other groups.

Performance for Test Data. As a result of the prediction of test data using the prediction model, the confusion matrix of Table 4 was obtained. Over 50% recall was found in all the groups, with over 50% precision found in all prediction groups, resulting in >65% accuracy for the overall data.

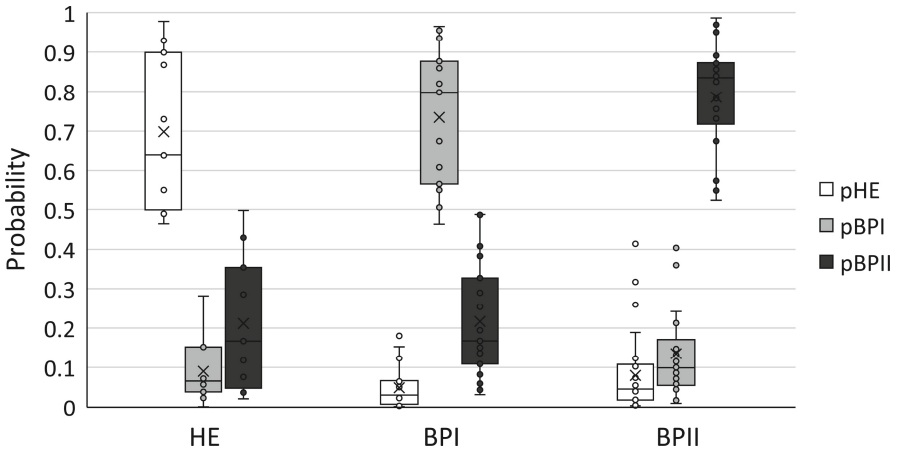

Fig. 1. The distribution of the group discrimination probability of the subjects used for training belonging to each group.

The distribution of the group discrimination probability of the subjects belonging to each group is presented in Fig. 2. In the HE group, the probability of the subjects being classified into the same group as the one they originally belonged to tended to be higher than the probability of the subjects being classified into other groups. Concerning the BPI and BPII groups, an overlap in distribution was observed between the probabilities of being classified into the BPI and BPII groups.

Table 4. The confusion matrix of test data using the prediction model.

		Prediction				
		HE	BPI	BPII	Recall	
Actual	HE	3	0	0	100%	
	BPI	1	3	2	50.0%	
	BPII	0	3	6	66.7%	
	Precision	75.0%	50.0%	75.0%	Accuracy	66.7%

4 Discussion

Concerning the prediction model performance, the overall accuracy was generally favorable for the training data. The accuracy for the test data was lower than that of the training data. Therefore, the prediction model was determined to be

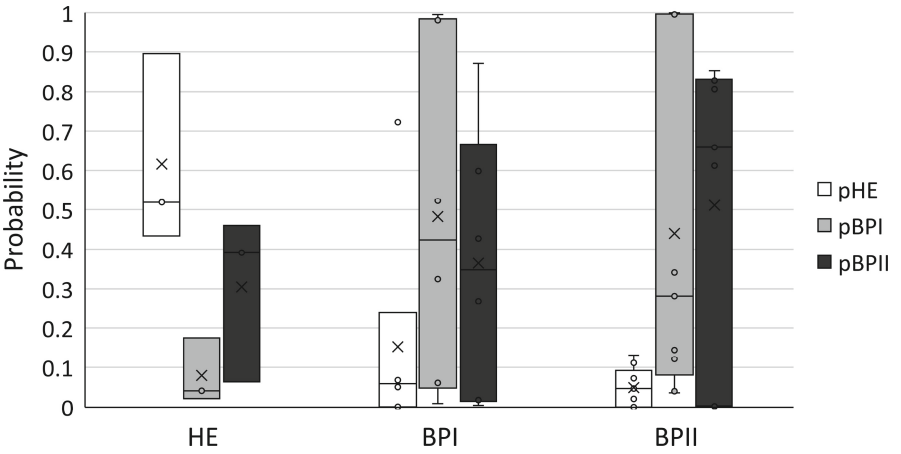

Fig. 2. The distribution of the group discrimination probability for testing subjects belonging in each group.

over-fitting, that is, a model optimized exclusively for learning data, and to be a model with inferior universality, that is, a model with low accuracy for data excluding the training data. Although the prediction model was over-fitting, it classified the training data, divided into three groups, with high accuracy, suggesting that the vocal features discriminate bipolar disorder. Moreover, if the BPI and BPII groups were merged, the prediction model may classify the healthy subjects and the patients with high accuracy.

Concerning the confusion matrix of the test data, the low recall in the BPI and BPII groups may be due to the similarities among the voices in the BPI and BPII patients. Even at the feature selection stage, the BPI and BPII groups could not be discriminated with an AUC of >0.8 for a single feature. Furthermore, this may indicate that the voices belonging to the two groups were similar; however, there were few subjects occasionally presenting equal probability of being classified into either of the BP groups. As shown in Tables 1 and 2, there was a difference in the numbers of males and females, age, HAM-D and YMRS score between the training and the test data. This can be attributed to the small sample size; moreover, it was impossible to match the training data and the test data sufficiently and may have affected the results.

In this study, audio data were collected in one setting. As such, the possibility of the data being impacted by the environment cannot be eradicated. In the future, it is necessary to collect audio data in other locations as well and to improve the prediction accuracy of the model.

In this study, analysis does not mention the details on the vocal features used in the model. Verifying which characteristics in the patient's voice were captured in the selected feature is another task for the future.

5 Conclusion

In this study, a vocal evaluation index was proposed to classify patients with bipolar type I or type II disorder by using patients with bipolar disorder as targets and healthy individuals as controls. By extracting vocal features from the voices of the subjects and selecting vocal features effective for a model, a polytomous logistic regression analysis was conducted to construct a prediction model for classifying healthy individuals and those with bipolar I or bipolar II. When the subjects in the test data were classified using the prediction model, the subjects were classified into three categories at an accuracy of approximately 67%. We suggest that the vocal index could be a new evaluation index for classifying bipolar disorders.

References

1. Izawa, S., et al.: Salivary dehydroepiandrosterone secretion in response to acute psychosocial stress and its correlations with biological and psychological changes. Biol. Psychol. **79**(3), 294–298 (2008)
2. Suzuki, G., et al.: Decreased plasma brain-derived neurotrophic factor and vascular endothelial growth factor concentrations during military training. PloS One **9**(2), e89455 (2014)
3. Garcia, R.G., Valenza, G., Tomaz, C.A., Barbieri, R.: Instantaneous bispectral analysis of heartbeat dynamics for the assessment of major depression. In: The Proceedings of Computing in Cardiology 2015, pp. 781–784. Nice (2015)
4. Goldberg, D.P.: Manual of the General Health Questionnaire. NFER Publishing, Windsor (1978)
5. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measureing depression. Arch. Gen. Psychiatry **4**(6), 561–571 (1961)
6. Young, R.C., Biggs, J.T., Ziegler, V.E., Meyer, D.A.: A rating scale for mania: reliability, validity and sensitivity. Br. J. Psychiatry **133**(5), 429–435 (1978)
7. Delgado-Rodriguez, M., Llorca, J.: Bias. J. Epidemiol. Community Health **58**(8), 635–641 (2004)
8. Cummins, N., Epps, J., Breakspear, M., Goecke, R.: An investigation of depressed speech detection: features and normalization. In: The Proceedings of the 12th Annual Conference of the International Speech Communication Association, Florence, pp. 2997–3000 (2011)
9. Mundt, J.C., Vogel, A.P., Feltner, D.E., Lenderking, W.R.: Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psychiatry **72**(7), 580–587 (2012)
10. Tokuno, S., Mitsuyoshi, S., Suzuki, G., Tsumatori, G.: Stress evaluation by voice: a novel stress evaluation technology. In: The Proceedings of the 9th International Conference on Early Psychosis, Tokyo, pp. 17–19 (2014)
11. Jiang, H., et al.: Investigation of different speech types and emotions for detecting depression using different classifiers. Speech Commun. **90**, 39–46 (2017)
12. Diagnostic and statistical manual of mental disorders V. American Psychiatric Association (2013)
13. Bowden, C.L.: Strategies to reduce misdiagnosis of bipolar depression. Psychiatr. Serv. **52**(1), 51–55 (2001)

14. Muzina, D.J., Kemp, D.E., McIntyre, R.S.: Differentiating bipolar disorders from major depressive disorders: treatment implications. Ann. Clin. Psychiatry **19**(4), 305–312 (2007)
15. Nakamura, M., et al.: Feasibility study of classifying major depressive disorder and bipolar disorders using voice features. In: The Proceedings of WPA XVII World Congress of Psychiatry, Berlin (2017)
16. Higuchi, M., et al.: Classification of bipolar disorder, major depressive disorder, and healthy state using voice. Asian J. Pharm. Clin. Res. **11**(3), 89–93 (2018)
17. Faurholt-Jepsen, M., et al.: Voice analysis as an objective state marker in bipolar disorder. Transl. Psychiatry **6**, e856 (2016)
18. Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., Morales, E.F.: Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive Mob. Comput. **31**, 50–66 (2016)
19. Sheehan, D.V., et al.: The mini-international neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry **59**(Suppl. 20), 22–33 (1998)
20. Hamilton, M.: A rating scale for depression. J. Neurol. Neurosurg. Psychiatry **23**, 56–62 (1960)
21. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE - the Munich versatile and fast open-source audio feature extractor. In: The Proceedings of the 18th ACM International Conference on Multimedia, Firenze, pp. 1459–1462 (2010)
22. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)
23. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/. Accessed 2 Dec 2018