Published in final edited form as:

Proc IEEE Int Conf Acoust Speech Signal Process. 2016 March ; 2016: 2359–2363. doi:10.1109/ICASSP.2016.7472099.

MOOD STATE PREDICTION FROM SPEECH OF VARYING ACOUSTIC QUALITY FOR INDIVIDUALS WITH BIPOLAR DISORDER

John Gideon[*], **Emily Mower Provost**[*], and **Melvin McInnis**[†]

[*]Department of Computer Science and Engineering, University of Michigan

[†]Department of Psychiatry, University of Michigan

Abstract

Speech contains patterns that can be altered by the mood of an individual. There is an increasing focus on automated and distributed methods to collect and monitor speech from large groups of patients suffering from mental health disorders. However, as the scope of these collections increases, the variability in the data also increases. This variability is due in part to the range in the quality of the devices, which in turn affects the quality of the recorded data, negatively impacting the accuracy of automatic assessment. It is necessary to mitigate variability effects in order to expand the impact of these technologies. This paper explores speech collected from phone recordings for analysis of mood in individuals with bipolar disorder. Two different phones with varying amounts of clipping, loudness, and noise are employed. We describe methodologies for use during preprocessing, feature extraction, and data modeling to correct these differences and make the devices more comparable. The results demonstrate that these pipeline modifications result in statistically significantly higher performance, which highlights the potential of distributed mental health systems.

Keywords

Bipolar Disorder; Mood Modeling; Mobile Health; Speech Analysis

1. INTRODUCTION

Bipolar disorder (BP) is characterized by swings in mood between mania, or heightened mood, and depression, or lowered mood. BP is pervasive, affecting 4% of people in the United States [1]. Both mania and depression profoundly impact the behavior of affected individuals, resulting in potentially devastating economic, social, and professional consequences. The current treatment paradigm involves routine monitoring of individuals through regular clinical visits. However, there are insufficient resources to ensure that all individuals with BP have access to this type of care [2]. This scarcity of available care points to the need for novel approaches to regular mood monitoring and the potential of computational approaches to serve as auxiliary methods. In this paper, we present an investigation into automatic speech analysis using mobile phone conversations as a way to predict mood, as well as the complications that arise from the diversity of real world recordings.

Research has demonstrated that speech patterns are affected by mood and contribute to accurate clinical assessments [3]. For example, both the Hamilton Depression Scale (HAMD) [4] and Young Mania Rating Scale (YMRS) [5] use clinical observations of speech to determine the severity of depression or mania [4, 5]. There is an opportunity to discover how speech cues can be automatically processed to augment objective measures available in clinical assessments. Mobile phones provide an effective platform for naturally monitoring these speech cues and have shown promise for BP [6, 7, 8]. However, changes in recording quality between different types of phones can severely decrease the predictive capabilities of a system. These include clipping, loudness, and background noise.

Much mood speech research has been centered around identifying speech features for recognizing depression. Among these, are pitch, energy, rhythm, and formants [9, 10, 11, 12, 13, 14]. Short pauses and increased pitch have been correlated with mania [10, 12, 14, 15, 16]. However, much of the work in identifying speech associated with mania has focused on differentiating it from schizophrenia and cannot be directly applied [17, 18]. Many mood related studies collected their speech from controlled environments [10, 12, 13] or used a single type of recording device [7, 8, 19] and do not necessarily reflect the variations in background noise and microphone quality present in real world recordings. As such, their models would be difficult to apply to a widely distributed mobile health system.

In this paper, we focus on one of the challenges associated with real-world distributed mood recognition: variability in recording. We examine the differences between the two phones used in this study and analyze preprocessing and modeling methods that allow us to build models of mood across the database as a whole. These methods include declipping [20], noise-robust segmentation [21], feature normalization [13], and multi-task learning [22]. We provide evidence that mood-related changes in speech are captured in this model using the structured assessment calls captured from different phone types. Please see Figure 1 for a system overview.

The novelty of our approach is the investigation into acoustic variations caused by recording with different types of phones and the preprocessing and modeling changes necessary to detect mood under these conditions. Our results suggest that this pipeline of methods including preprocessing, feature extraction, and data modeling can effectively increase the performance of these types of mixed device systems. The results show a significant increase in performance from AUCs of 0.57±0.25 and 0.64±0.14 for manic and depressed, respectively, to 0.72±0.20 and 0.75±0.14, highlighting the importance of proper processing of acoustic data from multiple sources.

2. PRIORI DATASET

The PRIORI Dataset is an ongoing collection of smartphone conversational data (reviewed and approved by the Institutional Review Board of the University of Michigan, HUM00052163). The participants are recruited from the HC Prechter Longitudinal Study of Bipolar Disorder at the University of Michigan [23]. The inclusion criteria are a diagnosis of rapid-cycling BP, type I or II. The exclusion criteria are a history of substance abuse and neurological illness.

Participants are enrolled for six to twelve months and are provided with an Android smartphone with the secure recording application (*PRIORI app*) installed. The app runs in the background and turns on whenever a phone call is made, recording only the participant's side of the dialog. The speech is encrypted in real-time, stored on the phone, and then uploaded to a HIPAA-compliant server.

2.1. Data Description

The recorded calls are designated into one of two groups: assessment and personal. Participants take part in weekly calls with our study clinicians in which the HAMD and YMRS interviews are conducted. The assessment calls establish a ground truth for the participant's mood over the previous week. The remainder of the data are referred to as personal calls. The personal calls represent all calls that take place outside of the clinical context. These calls are not annotated to ensure patient privacy and are not used in this study.

The PRIORI Bipolar Dataset currently contains 37 participants who have made 34,830 calls over 2,436 hours. Each participant has been on the study for an average of 29.2 weeks with a standard deviation of 16.4 weeks. Additionally, there have been 780 recorded weekly clinical assessments. Only these structured calls are used in this study. Twenty-three of these assessments were transcribed with speech and silence locations to aid in the determination of segmentation parameters.

2.2. Label Assignment

The HAMD and YMRS scales are continuous measures of mood, ranging from a score of 0 (not symptomatic) to 34 (highly symptomatic). In this paper, we treat the prediction problem as classification, binning the HAMD and YMRS into categories of symptomatic (depressed or manic, respectively) and asymptomatic (euthymic). Scores under a threshold of 6 on both scales are assigned a label of euthymic. Scores above 10 on the HAMD and below 6 on YMRS are assigned a label of depressed. Scores above 10 on the YMRS and below 6 on the HAMD are assigned a label of manic. Data in six-ten range on either scale and data with labels above 10 on both scales are excluded. Table 1 shows the class distribution.

The large standard deviations seen in Table 1 demonstrate the widely varying amounts of mood episodes between individuals with BP. Additionally, some individuals have disparities among the proportions of times spent in each mood. For example, one participant experienced 27 weeks of euthymia and two weeks of mania. Techniques to handle this imbalance are discussed in Section 5.

2.3. Phone Model Differences

The Samsung Galaxy series of phones, including the S3, S4, and S5 are used by participants. Only two of the participants were given S4s and their data are excluded from this study. The distribution of subjects with S3s and S5s can be seen in Table 2. The two models of phone include model-specific microphones and processing. One of the effects of this recording and processing is clipping. Clipping occurs most often in the S3, with an average of 2.74% of speech samples at maximum range. This sensitivity is also demonstrated by the average root mean square value of 0.397 for the S3. Additionally, the noise is much more pronounced, as seen in the lower signal to noise ratio of 21.2 dB for the S3.

3. PREPROCESSING

The two phones used in this study have different acoustic properties. The S3, compared to the S5, has more clipping, higher volume, and a sensitivity to background noise. Because of this, it is necessary to carefully preprocess the data before feature extraction using declipping, audio normalization, and noise-robust segmentation in order to make calls from different devices more comparable.

Declipping

The declipping algorithm *Regularlized Blind Amplitude Reconstruction* (RBAR) [20] was used to approximate the original signal. This is a closed form solution approximation of an algorithm called *Constrained Blind Amplitude Reconstruction* (CBAR) [24]. Each algorithm extrapolates the clipped sections of audio beyond their original values, while minimizing the second derivative of the signal, and have been shown to improve the performance of automatic speech recognition [20, 24]. Both algorithms ignore unclipped regions, beneficial for audio recordings that have variable amounts of clipping, as seen in Table 2.

Audio Normalization

The audio signal is scaled by dividing by the maximum absolute value. This ensures that the signal ranges from −1 to 1, which is necessary after running declipping, as it extrapolates the signal beyond these bounds. It also ensures that the loudness between the two phone types, as seen in Table 2, is more comparable.

Segmentation

Each call is segmented using an extension of Sadjadi and Hansen’s algorithm [21], which is robust to variation in noise. This is necessary, given the differences in SNR between the phones (Table 2). The algorithm extracts five signals representative of speech likelihood, including: harmonicity, clarity, prediction gain, periodicity, and perceptual spectral flux. These are then combined using principal component analysis (PCA). The final signal is the largest eigenvalue. It is smoothed by a Hanning window of 25ms and normalized by subtracting by the 5th percentile over the call and dividing by the standard deviation. This ensures that signals from different calls all share a similar silence baseline. Segments of 25ms are created wherever the combo signal exceeds a 1.8 threshold. Overlapping segments are merged and any silences less than 700ms are removed. These parameters were found by validating over the transcribed assessments for segment alignment. Segments are further divided into subsegments of 2s with 1s overlap. Segments less than 2s are discarded. Constant window sizes are used to ensure that variations in the features are not caused by changes in segment size [25]. The full segmentation process is shown in Figure 2.

4. FEATURE EXTRACTION

Rhythm Features

Individuals in manic or depressed episodes exhibit changes in the rhythm of their speech [26]. Rhythm features are calculated for each subsegment by first extracting the voicing envelope. The envelope is used to calculate the spectral power ratio and spectral centroid. The envelope is decomposed into two intrinsic mode functions (IMF) using empirical mode decomposition [27]. Tilsen and Arvaniti [25] empirically demonstrated that the extracted IMFs are reflective of syllable- and word-level fluctuations. The IMFs are used to extract five segment-level features: the power ratio between the two IMFs and the mean and standard deviation of the instantaneous frequencies associated with each IMF.

Call-Level Statistics

The seven rhythm features are transformed into call-level features by taking the mean, standard deviation, skewness, kurtosis, minimum, maximum, range, and 1st, 10th, 25th, 50th, 75th, 90th, and 99th percentiles of the subsegment measures. Additionally, the differences between the 50th and 25th, 75th and 50th, 75th and 25th, 90th and 10th, and 99th and 1st percentiles are included. This set is augmented with the percentage of the call that is above 10%, 25%, 50%, 75%, and 90% of the range. Finally, the call-level feature trend is captured by fitting a linear regression model to the features extracted over each segment (R^2, mean error, and mean squared error). This results in a total of 217 features.

Feature Normalization

Call-level features are Z-normalized either (1) globally, using the mean and standard deviation of all training data, or (2) by subject, using the mean and standard deviation of each subject's own data. Previous research has shown that normalization by subject can reduce the disparity between subject feature distributions caused by speaker differences and aid in the detection of mood [13]. This method may also help reduce some of the differences in subject feature distributions due to differences in phones.

5. DATA MODELING

The classification goal is to identify if a given call is (1) from a manic or euthymic episode or (2) from a depressed or euthymic episode. Subjects are only included in analysis if they have at least six total assessments in order to ensure enough data to process features by subject. Additionally, subjects must contain at least two euthymic calls and two manic/depressed calls. This ensures that there is enough data to measure test performance. With these restrictions, 15 subjects are used when considering mania (12 S3s and 3 S5s) and 18 subjects are used when considering depression (11 S3s and 7 S5s).

Support Vector Machines (SVM) [28] are used to classify the speech. SVMs learn a decision boundary between two classes of data with an explicit goal of identifying a boundary that maximally separates the two classes. The classifiers are implemented using both linear and radial basis function (RBF) kernels. Euthymic samples are given a weight equal to the number of manic/depressed samples divided by the number of euthymic samples. Manic/

depressed samples are given a weight of one. This ensures that there is no bias towards the mood with more samples by increasing the penalty for misclassification of minority labels. Multi-task SVMs [22] are also used for certain experiments. This algorithm weights the kernel function using a parameter *rho* in order to decrease the importance of data from a different task. In this case, the task is considered to be the phone type. On one extreme, rho can be selected to behave as a single-task SVM and consider the tasks to be equal. On the other extreme, the selected rho can consider the tasks to be completely independent.

The models are trained using leave-one-subject-out cross-validation, ensuring that there is no overlap between the speakers used to train and test the system. The model parameters include: kernel type (RBF vs. linear), gamma (RBF only), number of features with respect to a ranked list, cost parameter (C), and rho (multi-task only). The parameter combination is chosen to optimize leave-one-training-subject-out cross-validation, where the contribution of each training subject is proportional to his/her amount of data.

Features are ranked using a heuristic of Weighted Information Gain (WIG). The heuristic was chosen due to the observed subject-specific label imbalance, which may result in the identification of features that are tied to subject identity, rather than mood. This can result in a classifier learning to associate all instances with a single mood state from a biased subject. WIG allows for each sample to be ascribed an importance that ensures both classes contribute equally from each subject. This is implemented using the weighted entropy functions described in [29]. Each sample is given a weight equal to the total number of samples in its subject divided by the number of occurrences of its label in its subject. This ensures that minority and majority samples are given equal weight over each subject, while subjects are given weight proportional to their number of samples.

The system performance was measured using Area Under the Receiver Operating Characteristic Curve (AUC). AUC assesses the ability of a system to correctly rank pairs of instances from opposing classes. It has a chance rating of 0.5 and ideal rating of 1.

6. RESULTS AND DISCUSSION

In this section we demonstrate the ability to differentiate between euthymic and symptomatic moods, despite using two types of mobile phones with different acoustics. The results are presented in Table 3. In addition to reporting the combined test performance of both phone types, results are broken down into individual types. However, all phones from both types are always used to train models. A paired t-test with a significance of 0.05 is used to compare results to baseline performance and a significant difference is marked with an asterisk and bolded.

Baseline Performance

The baseline system uses global normalization and does not include declipping. The results in Table 3a show an AUC of 0.64±0.14 for depressed and a near chance performance of 0.57±0.25 AUC for manic. However, the three S5s performed better than the S3s in the manic test with 0.78±0.31 AUC. This could indicate that even though the S5 only makes up 20% of the phones, its higher quality recordings allow for it to perform well in testing.

Alternately, the speaker population that makes up those subjects using the S5s could be more homogeneous. The S5 continues to outperform the S3 in the rest of the manic experiments.

Evaluation of Declipping

Table 3b shows the results of declipping when using global normalization. While the performance of the depressed tests remain mostly unaffected, the manic test increases significantly to an AUC of 0.70±0.17. This is due to the improvement in the S3, where larger amounts of clipping occurred, as seen in Table 2. We hypothesize that the stronger improvement in manic tests, compared with depressed tests, is due to the fact that manic S3 calls have significantly more clipping than euthymic and depressed S3 calls (unpaired t-test, p=0.05). The percent of clipping in euthymic, manic, and depressed S3 calls are 2.73±1.25%, 3.21±1.13%, and 2.41±1.07%, respectively.

Evaluation of Segmentation

The effect of segmentation was studied by eliminating the algorithm described in Section 3. Instead, the 2 second subsegments were taken over the entire call - silences included. It performed the best of all tests with significant increases from the baselines for both moods (Table 3c). However, we hypothesize that this is actually due to the rhythm features indirectly capturing information about the assessment structure. For example, an individual who is euthymic would have more silence due to their brief interview answers. This highlights one of the potential pitfalls to avoid when working with structured calls to train a model to recognize acoustic aspects of mood. For this reason, it is necessary to use accurate segmentation to avoid these misleading results.

Evaluation of Feature Normalization

Normalization by subject significantly increased the performance of both manic and depressed tests from baseline, as shown in Table 3d. This method has the ability to correct for different feature distributions among speakers, as explained in [13]. These results demonstrate that this correction can also benefit systems with variable recording devices of different quality.

Multi-task SVM Analysis

The use of a multi-task SVM can also control for the variability in device types by giving lower weight to data from different phone types and higher weight to data from the same phone types. Table 3e shows a significant improvement in manic from baseline by selecting a low value for rho and treating data from across different phone types as less informative. Depression does not see much improvement, as a high rho value is selected, indicating that the data is already comparable without preprocessing. This gives further evidence to the reason preprocessing works well for manic speech but has little effect on depressed speech. Another multi-task experiment was run using the preprocessing methods that worked best for each mood - RBAR declipping and subject normalization for manic and subject normalization for depressed. These results can be seen in Table 3f, with the highest manic AUC of 0.72±0.20, which is significantly better than baseline.

7. CONCLUSION

This paper presents methods to improve the comparability of data collected from across devices of different acoustics. This is essential for any mobile health system using speech that aims to be widely distributed, as the prospect of varying audio quality is unavoidable. Our results demonstrate that through certain preprocessing, feature extraction, and data modeling techniques it is possible to mitigate the effects of differing amounts of clipping, loudness, and noise. This is best shown by the increase in performance from the baseline AUCs of 0.57±0.25 for manic and 0.64±0.14 for depressed to the significantly higher AUCs of 0.72±0.20 and 0.75±0.14, respectively. This excludes the results without segmentation, as those features capture the structure of the mood interview instead of the characteristics of the speech. There was not a comprehensive solution for both mood types, which indicates the need for careful consideration of all steps along any pipeline.

The ultimate goal will be for the system to be totally passive, requiring no active input from the BP patient or the clinic. Current methods using structured assessments are not enough, as they require weekly interview calls. However, the transition to personal calls will require solutions to many problems, including how to control for the confounding factors of variations in subject symptomatology, episode patterns, and conversational styles. The refinement of techniques developed in this study to increase device comparability may be adaptable to these issues. In particular, it will be necessary to determine how to adapt the system to particular individuals and determine which features are indicative of mood and not some other misleading factor. Although, if effective, it will greatly assist in the way that mental health care is managed.

8. REFERENCES

[1]. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. Archives of general psychiatry. 2005; 62(no. 6):593–602. [PubMed: 15939837]

[2]. Angst J, Sellaro R, Angst F. Long-term outcome and mortality of treated versus untreated bipolar and depressed patients: a preliminary report. International Journal of Psychiatry in Clinical Practice. 1998; 2(no. 2):115–119. [PubMed: 24946291]

[3]. National Institute of Mental Health. [Accessed: September - 2015] Bipolar disorder in adults. http://www.nimh.nih.gov/health/publications/bipolar-disorder-in-adults/Bipolar_Disorder_Adults_CL508_144295.pdf

[4]. Hamilton, M. ECDEU Assessment Manual For Psychopharmacology, Revised Edition. National Institute of Mental Health; Rockville, MD: 1976. Hamilton depression scale; p. 179-92.

[5]. Young R, Biggs J, Ziegler V, Meyer D. A rating scale for mania: reliability, validity and sensitivity. The British Journal of Psychiatry. 1978; 133(no. 5):429–435. [PubMed: 728692]

[6]. Karam, Z.; Provost, E. Mower; Singh, S.; Montgomery, J.; Archer, C.; Harrington, G.; Mcinnis, M. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 2014; p. 4858-4862.

[7]. LiKamWa, R.; Liu, Y.; Lane, ND.; Zhong, L. Moodscope: building a mood sensor from smartphone usage patterns. ACM International Conference on Mobile Systems, Applications, and Services; 2013; ACM; p. 389-402.

[8]. Osmani, V.; Maxhuni, A.; Grünerbl, A.; Lukowicz, P.; Haring, C.; Mayora, O. Monitoring activity of patients with bipolar disorder using smart phones. Proceedings of the International Conference on Advances in Mobile Computing & Multimedia; 2013; ACM; p. 85

[9]. France D, Shiavi R, Silverman S, Silverman M, Wilkes DW. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Transactions on Biomedical Engineering. 2000; 47(no. 7):829–837. [PubMed: 10916253]

[10]. Vanello, N.; Guidi, A.; Gentili, C.; Werner, S.; Bertschy, G.; Valenza, G.; Lanata, A.; Scilingo, EP. Speech analysis for mood state characterization in bipolar patients. Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2012; p. 2104-2107.

[11]. Quatieri, TF.; Malyska, N. Vocal-source biomarkers for depression: A link to psychomotor activity. Proceedings from the Conference of the International Speech Communication Association (INTERSPEECH); 2012;

[12]. Guidi A, Vanello N, Bertschy G, Gentili C, Landini L, Scilingo E. Automatic analysis of speech f0 contour for the characterization of mood changes in bipolar patients. Biomedical Signal Processing and Control. 2014

[13]. Cummins, N.; Epps, J.; Breakspear, M.; Goecke, R. An investigation of depressed speech detection: Features and normalization. Proceedings from the Conference of the International Speech Communication Association (INTERSPEECH); 2011; p. 2997-3000.

[14]. Friedman E, Sanders G. Speech timing of mood dis orders. Computers in Human Services. 1991; 8(no. 3-4):121–142.

[15]. Resnick H, Oltmanns T. Hesitation patterns in the speech of thought-disordered schizophrenic and manic patients. Journal of Abnormal Psychology. 1984; 93(no. 1):80. [PubMed: 6699277]

[16]. Pogue-Geile M, Oltmanns T. Sentence perception and distractibility in schizophrenic, manic, and depressed patients. Journal of Abnormal Psychology. 1980; 89(no. 2):115. [PubMed: 7365124]

[17]. Taylor M, Reed R, Berenbaum S. Patterns of speech disorders in schizophrenia and mania. The Journal of Nervous and Mental Disease. 1994; 182(no. 6):319–326. [PubMed: 8201303]

[18]. Hoffman R, Stopek S, Andreasen N. A comparative study of manic vs schizophrenic speech disorganization. Archives of General Psychiatry. 1986; 43(no. 9):831–838. [PubMed: 3753163]

[19]. Grunerbl A, Muaremi A, Osmani V, Bahle G, Ohler S, Tröster G, Mayora O, Haring C, Lukowicz P. Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE Journal of Biomedical and Health Informatics. 2015; 19(no. 1):140–148. [PubMed: 25073181]

[20]. Harvilla, MJ.; Stern, RM. Efficient audio declipping using regularized least squares. Proceedings from the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); April 2015; p. 221-225.

[21]. Sadjadi SO, Hansen J. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Processing Letters. 2013; 20(no. 3):197–200.

[22]. Evgeniou, T.; Pontil, M. Regularized multi–task learning. ACM International Conference on Knowledge Discovery and Data Mining (KDD); 2004; ACM; p. 109-117.

[23]. Langenecker SA, Saunders E, Kade AM, Ransom MT, McInnis MG. Intermediate: cognitive phenotypes in bipolar disorder. Journal of Affective Disorders. 2010; 122(no. 3):285–293. [PubMed: 19800130]

[24]. Harvilla, MJ.; Stern, RM. Least squares signal declipping for robust speech recognition. Proceedings from the Conference of the International Speech Communication Association (INTERSPEECH); 2014;

[25]. Tilsen S, Arvaniti A. Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. The Journal of the Acoustical Society of America. 2013; 134(no. 1):628–639. [PubMed: 23862837]

[26]. Goodwin, FK.; Jamison, KR. Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression. Oxford University Press; 2007.

[27]. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen N, Tung CC, Liu HH. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences. 1998; 454(no. 1971):903–995.

[28]. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(no. 3):273–297.

[29]. Śmieja M. Weighted approach to general entropy function. IMA Journal of Mathematical Control and Information. 2014:dnt044.

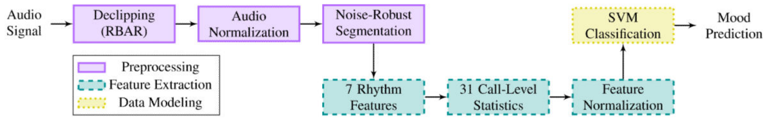

Fig. 1.

Audio pipeline divided into three stages of preprocessing (Section 3), feature extraction (Section 4), and data modeling (Section 5).

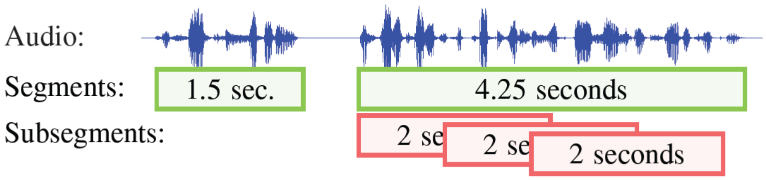

Fig. 2.

Segments of speech are found. Segments of 2 seconds or longer are divided into subsegments of 2 seconds in 1 second steps.

Table 1

Distribution of assessment classes of mood. The total number of observations of each mood class is given. The mean and standard deviation of observations for each class per subject is shown, along with the average percentage of each.

Mood	Total	# Per Subject	% Per Subject
Euthymic	275	7.9±7.7	30%
Manic	107	3.1±4.0	12%
Depressed	247	7.1±7.5	28%
Mixed	95	2.7±3.6	13%
Excluded	175	5.0±4.7	17%

Table 2

Differences in data amounts and acoustics between the Galaxy S3 and S5. The percent clipped assessments (Assess.) and the mean percent of samples per call clipped are shown. Root mean square (RMS) values are calculated to show the loudness for each device microphone. Signal to noise ratio (SNR) is calculated as the relative power in the speech verses silence regions in decibels (dB).

Phone	#Subjects	#Assess.	%Clipped	RMS	SNR_{dB}
S3	18	456	2.74%	0.397	21.2
S5	17	287	0.02%	0.066	25.1
Both	35	743	1.69%	0.269	23.1

Table 3

Classification results using various methods. **Bolded*** AUCs denote results significantly better than baseline (paired t-test, p=0.05).

Model	Manic AUC	Depressed AUC
S3	0.52±0.22	0.66±0.17
S5	0.78±0.31	0.62±0.09
Both	0.57±0.25	0.64±0.14

(a) No Declipping and Global Normalization (Baseline)

Model	Manic AUC	Depressed AUC
S3	0.68±0.16	0.62±0.14
S5	0.79±0.21	0.69±0.18
Both	**0.70±0.17***	0.65±0.15

(b) RBAR Declipping and Global Normalization

Model	Manic AUC	Depressed AUC
S3	0.73±0.22	0.74±0.10
S5	0.79±0.37	0.80±0.21
Both	**0.74±0.24***	**0.77±0.15***

(c) No Speech Segmentation (Silence Included)

Model	Manic AUC	Depressed AUC
S3	0.66±0.15	0.73±0.15
S5	0.71±0.35	0.78±0.10
Both	**0.67±0.19***	**0.75±0.14***

(d) No Declipping and Subject Normalization

Model	Manic AUC	Depressed AUC
S3	0.67±0.20	0.67±0.21
S5	0.72±0.41	0.65±0.11
Both	**0.68±0.23***	0.66±0.18

(e) Multi-Task SVM Using Baseline Preprocessing

Model	Manic AUC	Depressed AUC
S3	0.71±0.19	0.66±0.14
S5	0.78±0.23	0.79±0.13
Both	**0.72±0.20***	0.71±0.15

(f) Multi-Task SVM Using Best Preprocessing