Recognition of Depression in Bipolar Disorder: Leveraging Cohort and Person-Specific Knowledge

Soheil Khorram[⋆†]*, John Gideon*[⋆]*, Melvin McInnis*[†]*, Emily Mower Provost*[⋆]

Departments of: Computer Science and Engineering[⋆] and Psychiatry[†], University of Michigan

{khorrams, gideonjn, mmcinnis, emilykmp}@umich.edu

Abstract

Individuals with bipolar disorder typically exhibit changes in the acoustics of their speech. Mobile health systems seek to model these changes to automatically detect and correctly identify current states in an individual and to ultimately predict impending mood episodes. We have developed a program, PRIORI (Predicting Individual Outcomes for Rapid Intervention), that analyzes acoustics of speech as predictors of mood states from mobile smartphone data. Mood prediction systems generally assume that the symptomatology of an individual can be modeled using patterns common in a cohort population due to limitations in the size of available datasets. However, individuals are unique. This paper explores person-level systems that can be developed from the current PRIORI database of an extensive and longitudinal collection composed of two subsets: a smaller labeled portion and a larger unlabeled portion. The person-level system employs the unlabeled portion to extract i-vectors, which characterize single individuals. The labeled portion is then used to train person-level and population-level supervised classifiers, operating on the i-vectors and on speech rhythm statistics, respectively. The unification of these two approaches results in a significant improvement over the base-line system, demonstrating the importance of a multi-level approach to capturing depression symptomatology.

Index Terms: Bipolar Disorder, Depression Recognition, i-vectors, Mobile Health, Hybrid Classification

1. Introduction

Bipolar disorder (BP) is a severe and chronic illness characterized by pathological swings in mood, ranging from mania to depression [1]. It affects 2.6% of American adults [2] and has devastating effects on an individual's life, family, and work [3]. It is among the top 10 leading causes of disability in the United States [4] with up to 20% of people affected taking their own life [1]. These negative effects increase the necessity and urgency of monitoring and prioritizing care to mitigate serious episodes. To this end, mobile health technology can be useful for longitudinal monitoring the health of individuals [5–7]. In this paper, we use speech gathered from mobile phone conversations during the *PRIORI* project [8, 9] to train models for predicting depression in patients with BP. The proposed techniques exploit both *population-general and subject-specific* knowledge.

Speech is modulated by the mood of an individual [10]. In particular, the Hamilton Depression Rating Scale (HAMD) [11], lists speech retardation as one of the indicators of depression. Previous work showed that it is possible to augment the clinical diagnosis of depression with objective rating by automatic detection from speech [12–14]. This could be used to help to better target care to those most in need and help in areas with scarce resources [15]. However, variations in individual symptoms make the adoption of such a system difficult.

Many computational models have been proposed to predict depression from speech [6, 12–14, 16–20]. These models are normally trained to capture common patterns in cohorts due to limitations in the size of available datasets. For example, [16] explores the performance of common speaker-independent classifiers for detection of depression. Additionally, [17] compares Guassian Mixture Model (GMM) and Support Vector Machine (SVM) classifiers using only formant frequencies and their dynamics. Subject-independent systems using i-vectors have recently been proposed for depression detection [14, 18–20]. One important issue of using i-vectors for depression detection is the limited dataset available [21]. [14] proposes an oversampling approach to increase the number of available utterances.

There are a limited number of systems that leverage speaker-specific information in their classifiers. The work reported in [6] incorporates speech patterns along with other sensor modalities collected from smartphone devices to predict mood using a subject-specific classifier. Additionally, the work by Vanello's group [12, 13] found that jitter, pitch, and pitch contours were effective indicators of mood from a subject-specific perspective. However, in order to detect mood effectively on a large scale it is necessary to both understand population level indicators in addition to subject-specific variations.

The PRIORI database is a longitudinal collection of cellphone speech data from individuals with BP [8, 9]. It contains a considerable amount of data for each participant. This enables us to capture subject-specific as well as population-general aspects of speech. In this paper, we use an SVM trained with rhythm to characterize population-general speech [9]. We propose the use of *i-vectors* [22] extracted over *Mel-Frequency Cepstral Coeffiients (MFCC)* to capture mood variation at the subject-level using a *speaker-dependent SVM*. We leverage the large unlabeled subset of our data to circumvent the sparsity problems that often accompany i-vector extraction. Further, we employ the *Within-Class Covariance Normalization (WCCN)* technique [23] on the total variability subspace to alleviate the undesirable effect of mobile phone channels.

The main novelty of our approach is our fusion of a subject-specific system using unlabeled personal calls with a population-general system for the detection of depression from BP speech. We compared this fusion with the baseline system from [9], which modeled rhythm features in a population-general manner. Our results showed significant improvement from the baseline with the Unweighted Average Recall (UAR) increasing from 0.66 ± 0.11 to 0.73 ± 0.09 and the Area Under the receiver operating Curve (AUC) increasing from 0.69 ± 0.15 to 0.78 ± 0.12. This shows the importance of using both cohort and subject-specific knowledge when modeling mood.

Mood	Total	# Per Subject	% Per Subject
Euthymic	306	7.1±6.5	36%
Depressed	266	6.2±6.8	27%
Excluded	361	8.4±7.0	37%

Table 1: *Distribution of mood in the assessments. Shown are the total number of observations, the mean and standard deviation of subject observations, and the mean percentage of each.*

2. PRIORI Dataset

The PRIORI dataset is a large-scale collection of smartphone conversational speech from individuals with BP. The inclusion criteria are a diagnosis of BP type I or II and the exclusion criteria are a history of substance abuse and/or co-morbid neurological illness. Participants are enrolled for 6-12 months and are given a smartphone with a secure recording application installed, which records their side of every call made. The PRIORI dataset is actively growing and currently contains data from 43 participants with an average collection duration of 21.2±14.2 weeks per subject, including 39,445 calls and over 2,880 hours of speech. All experiments, including the baseline, are performed on the dataset snapshot at this time.

Participant mood is assessed weekly, over the phone, by a member of the study team using the Hamilton Depression Scale (HAMD) [11] and the Young Mania Rating Scale (YMRS) [24]. These calls are referred to as *assessment calls*. The dataset includes 933 recorded weekly clinical assessments, 23 of which were transcribed for algorithm development. All other calls are referred to as *personal calls*. These calls are neither annotated nor transcribed to protect patient privacy. The data include phone calls collected only when an individual was not using the speaker phone. See [9] for more details.

This paper focuses on depression. The labels are binned into three categories: euthymic, depressed, and excluded. Euthymic is defined as a score of six or less on both the HAMD and YMRS scales. Depressed is defined as a score of ten or greater on the HAMD and less than ten on the YMRS, as in [9]. The goal of this paper is to differentiate between euthymic and depressed speech. The speech in the excluded class is not considered. See Table 1 for the data distribution.

3. Feature Extraction

3.1. Rhythm

Rhythm features have been shown effective in detecting mood in cohorts in previous work [9]. In particular, depressed individuals exhibit speech that is slowed and has frequent pauses [1]. Only assessment calls are used in rhythm extraction.

Segmentation: Calls are segmented using a noise-robust method by Sadjadi and Hansen [25] which compensates for variations in background noise. Their algorithm extracts five representations of speech likelihood including: harmonicity, clarity, prediction gain, periodicity, and perceptral spectral flux. Principal Component Analysis (PCA) is performed to combine them into a single signal by taking the largest eigenvalue. We extend this approach by first converting this signal into contiguous speech segments. We then smooth the signal with a Hanning window of 25ms and normalize it by subtracting by the 5th percentile over the call and dividing by the standard deviation to ensure comparability between calls. Segments of 25ms are created whenever this signal exceeds a 1.8 threshold. This forms a set of overlapping segments which are merged, removing any silence less than 700ms. We determined these parameters by validating over the transcribed assessments. Segments longer than 2s are further divided into subsegments of 2s with 1s overlap. Constant segment size is used to ensure that variations in features are only due to variations in rhythm [9].

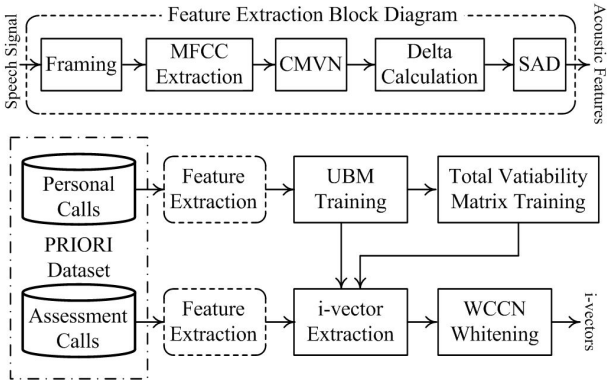

Figure 1: *Schematic block diagram of i-vector extraction.*

Rhythm Features: The rhythm features are calculated for each subsegment using an algorithm by Tilsen and Arvaniti [26]. The audio envelope is extracted and the spectral power ratio and centroid are found. The first two Intrinsic Mode Functions (IMF) are then extracted using empirical mode decomposition [27]. The power ratio between the IMFs, as well as the mean and standard deviation of their instantaneous frequencies are found. This forms a total of seven segment-level statistics that have shown to be related to syllable- and word-level rhythm [26]. A total of 31 statistics are used to form the call level feature vector of 217 dimensions, as in [9]. These include mean, standard deviation, skewness, kurtosis, minimum, maximum, and range. Additionally, we calculated various percentiles and percentile ranges, the percentage of the call above thresholds of the range, and linear regression coefficients and error.

3.2. i-vectors

Subject-specific mood variation is captured using i-vector representation. Recent works have demonstrated the efficacy of this technique for predicting depression over a cohort [14, 18–20].

i-vector Formulation: Speech contains many sources of variability, including identity [22], age [28], gender [28], and critically, mood [14]. These variations can be captured using the i-vector technique. Its underlying assumption is that factors of variation lie in a low-dimensional subspace spanned by the columns of the total variability matrix T, a low-rank rectangular matrix. An arbitrary speech instance, u, can be represented by a *GMM mean supervector*, $M(u)$, which is modeled by:

$$M(u) = m + Tw(u) \tag{1}$$

where m is the supervector constructed from the *Universal Background Model (UBM)* trained using all *personal call* data, $w(u) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an utterance-dependent identity vector (i-vector) [22]. The total variability matrix is trained using an *Expectation Maximization (EM)* algorithm introduced in [29].

Acoustic Features: 19 MFCCs and log energy are extracted using a 25ms Hamming window with 10ms step from the calls with longer than five seconds of speech. They are normalized using utterance-level *Cepstral Mean and Variance Normalization (CMVN)* [30] to compensate for background and channel noise. This 20-dimensional feature vector is applied to a feature warping [31] with 3 second sliding window. The final feature set contains the MFCCs/log energy, their Δ, and $\Delta\Delta$.

i-vector Extraction: Figure 1 shows the system developed for extracting i-vectors. We train the UBM (2048 Gaussians) and total variability matrix (400 dimensions) using the acoustic features. Then we extract assessment i-vectors and apply WCCN [23] on them to compensate for residual channel effect.

3.3. Feature Normalization:

Both feature sets are normalized using the mean and standard deviation of each subject. Additionally, each fold is globally normalized so that a mean of zero and standard deviation of one is attained across all subjects.

4. Data Modeling

SVMs [32] are used to classify both types of features. SVMs find the boundary that maximally separates two classes. We use either a linear or Radial Basis Function (RBF) kernel. We weight the samples to accommodate for class imbalance. Finally, the output score is the signed distance to the hyperplane.

Various divisions between training, testing, and validation folds are used in this paper and are defined below:

- **Population-General Validation:** One test subject is left out when training the model. This builds a system that is generalized to work on previously unseen individuals. We validate parameters by dividing the training subjects into four folds. We require that subjects have at least six calls, including at least two euthymic and two depressed, to ensure enough data for normalization and performance metric calculation. This *baseline system* was presented in our prior work [9].
- **Subject-Specific Validation:** Only data from one subject is used. During training one test call is left out. This produces a system that can adapt to the features of an individual. Validation is performed over the training calls divided into ten folds. Only subjects with at least four euthymic and four depressed calls are used to ensure enough training data. This system is used for i-vectors because they are suited to subject-specific modeling, as explained in Section 6.
- **Hybrid Validation:** A hybrid approach of the above two systems. Calls across all subjects are used. We train the model by leaving one test call out. This system uses information from both the subject of the test call and the population. Validation is performed over training calls divided between 10 folds with calls from all subjects distributed across folds. Only subjects with at least six calls, including at least two euthymic and two depressed, are used.

The kernel type and parameters (cost, gamma) as well as feature set size are selected during validation by maximizing *UAR* - the mean percentage of each class correctly identified. UAR gives minority classes equal weight. Features are ranked using a heuristic of Weighted Information Gain (WIG) to correct for subject label imbalances. This is implemented by weighted entropy as described in [33]. Each sample weight is set to the number of subject calls divided by the number of calls in the subject with the same label. This ensures that the sum of subject weights is proportional to its total number of samples, while also giving minority and majority labels equal weight. Only the test performances of subjects used in all systems are reported to make system results comparable.

To find the effect of combining cohort and person-specific knowledge, four fusion methods are considered (Figure 2):

Feature Fusion: Rhythm and i-vectors are concatenated into one feature vector. Hybrid validation is then performed.

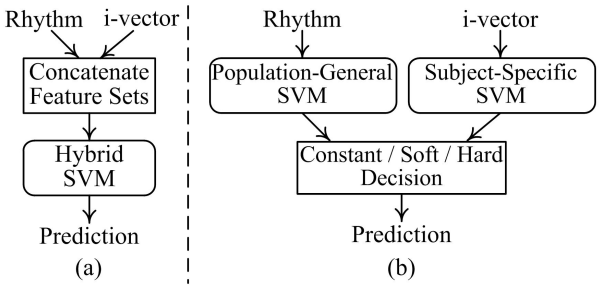

Figure 2: *Diagrams of the system fusions. (a) Hybrid modeling with concatenated features. (b) Constant, soft, and hard decision fusions (all in one figure).*

Decision Fusion: We train a rhythm population-general model and an i-vector subject-specific model. SVM outputs from both models are normalized using a sigmoid to ensure they are comparable. We determine the ideal weight (λ) to combine these systems for each test call using subject-specific validation. We find the population-general scores (PG) for the system trained using all but the test subject. We determine the subject-specific scores (SS) through validation by leaving one additional call out. This is necessary because test subject data is used in the model. Equation 2 shows the fusion scores (F):

$$F = PG \times \lambda + SS \times (1 - \lambda) \quad (2)$$

A higher weight indicates a higher contribution from the population-general system. Because score fusion is performed per call, each test call will have a different λ. During validation, we determine the weight based on the following three methods:

- **Constant:** λ is set to 0.5.
- **Soft:** λ is chosen between 0% and 100% by validating over increments of 1%. The best λ is selected by maximizing *UAR* measure. For subjects with fewer scores, there are often many weights that achieve the maximum UAR performance. A tie-breaking heuristic is used to determine which weight to choose in this case. The largest contiguous range of weights producing the maximum performance is found and the center weight is selected as λ. This mechanism was chosen to increase the stability of the fusion by selected a weight that is furthest from other weights that cause drops in performance.
- **Hard:** Same as soft, except λ is only allowed to be 0% or 100%. This results in only one classifier being selected.

5. Results

In addition to *UAR*, *AUC* is used in testing to compare the systems. *AUC* determines the system's ability to relatively rank test outputs. Unlike UAR, it does not have a set threshold. It is calculated as the area under the curve defined by the amount of true positives and false positives at all possible thresholds. Both measures have a chance performance of 0.5 and an ideal performance of 1. Table 2 shows a summary of results, including the two component systems and four fusions of systems.

Component Systems: The baseline system from [9] generalizes rhythmic symptoms across individuals. The subject-specific i-vector system learns individual patterns in the voice. Between the two, rhythm has the lower standard deviation of 0.11 UAR. Additionally, as seen in Table 3, no subjects perform worse than chance when using the rhythm model, showing greater stability. We hypothesize that this stability is due to the relatively larger number of samples across subjects used to

System	UAR	AUC
Population-General (Rhythm)	.66±.11	.69±.15
Subject-Specific (i-vector)	.64±.17	.70±.18
Feature Fusion	.71±.14	**.76±.13***
Constant Decision Fusion	.72±.15	.74±.16
Soft Decision Fusion	**.73±.09***	**.78±.12***
Hard Decision Fusion	.71±.11	.76±.13

Table 2: *Results for different systems (top) and fusions (bottom). Stared and bolded results mark significantly better performance than population-general baseline. (pairwise t-test, $p<0.05$).*

train the population-general model. This means that the model remains mostly consistent between test subjects. Contrast this with the subject-specific i-vector system where the entire model changes between subjects. This results in a standard deviation of 0.17 UAR (Table 2) and four subjects performing below chance (Table 3). However, the i-vector model performs better than the rhythm model for six subjects. This difference in performance can be further quantified by the low correlation of 0.12 between SVM outputs of the two component systems. Prior work has shown that uncorrelated systems are more effectively fused [34, 35]. This indicates that fusing population-general rhythm and subject-specific i-vectors will likely produce better results than either of its components.

Fusion: Soft decision fusion attained the best performance of all experiments with a significant (paired t-test) improvement over baseline of 0.73±0.09 UAR (p=0.045) and 0.78±0.12 AUC (p=0.020). We hypothesize that soft decision fusion works best, because it allows for direct tuning of subject contributions from both features and validation methodologies. There are strong correlations of 0.70 and -0.67 between the respective rhythm and i-vector UARs and the mean selected weight. This demonstrates its effectiveness at selecting the best performing component system, unlike the constant decision version which is unable to moderate their contributions and has less consistent subject results (not significant, p=0.14). Additionally, there are four instances where the soft decision fusion performs better than both component systems (highlighted in Table 3). This occurs when the weight is near 50% and there is contribution from both systems. This demonstrates its main advantage over the hard select technique - hard select can generally only perform as good as the best component system for each subject.

6. Discussion

The spatial distribution of the extracted i-vectors provides insight into the effectiveness of this feature at both the individual- and population-level. The i-vectors are mapped to a two-dimensional space using t-Distributed Stochastic Neighbour Embedding (t-SNE) [36], a dimensionality reduction algorithm that maps similar objects to nearby points and dissimilar objects to distant points. Figure 3 shows the distribution of the euthymic and depressed calls (blue and red dots in the figure, respectively). The individual groupings of the calls are at the subject-level. This separation at the subject-level suggests that the technique is effective for differentiating between mood within a speaker, but not between speakers due to the large speaker-effect. This speaker-effect would preclude the use of a population-general i-vector classification. This effect can be mitigated using subject-specific normalization. However, this creates strong overlap between the two mood categories, suggesting that i-vectors may be most effective at the individual, rather than population, level.

Rhythm	i-vector	Fusion	Mean λ	#Eut.	#Dep.	#Per.
.544	.730	.730	.10	9	14	474
.549	.761	.752	.18	19	14	513
.531	.813	.698	.22	6	16	2832
.707	.829	.829	.30	7	10	327
.570	.750	.740	.47	25	10	1660
.757	.714	.786	.56	5	7	780
.519	.385	.596	.57	13	4	348
.607	.400	.636	.62	7	20	769
.762	.774	.690	.63	7	6	131
.769	.625	.923	.64	26	4	1382
.714	.477	.618	.82	11	10	814
.739	.578	.683	.93	10	9	1483
.833	.417	.792	.93	6	12	558
.66±.11	.64±.17	.73±.09	.54±.27	12±7	10±5	929±741

Table 3: *Subject AUCs of Soft decision fusion and component systems, ordered by mean weight (λ). The number of euthymic (Eut), depressed (Dep), and personal (Per) calls are shown. The last row is the column means and standard deviations. Highlighted rows show when soft decision performs best.*

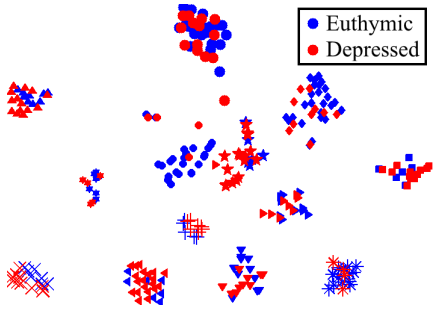

Figure 3: *t-SNE plot of i-vectors showing subject separability. Shapes represent subjects, while the colors depict moods.*

7. Conclusion

This paper demonstrates the importance of capturing both cohort and subject-specific variations in speech to effectively detect depression in bipolar disorder. This is important for a mental health monitoring system that both aims to be able to provide immediate help to new users and improved performance over time. Our paper introduces a soft decision fusion of population-general rhythm detection and subject-specific i-vector variation monitoring. The i-vectors are trained in a novel manner by using the unlabeled personal calls of individuals to learn patterns in the speech of subjects. This allows for a subject-specific system to be trained with relatively few assessment calls to effectively model individual changes in depression. The results show that the fusion significantly improves performance from the baseline of population-general rhythm.

The fusion experiments concentrated on learning the weights in a subject-specific manner because at least 4 samples of each label type were available. While this type of system worked well for depression, it may be difficult to adapt to mania. On average, subjects tend to be depressed three times more often than manic [37]. Due to this lack of data it may be necessary to consider similar subjects with more data as part of the fusion. However, this would require an investigation into subject similarities between speech phenotypes and mood. This is becoming increasingly important as we begin to model the personal calls, as population-general symptoms may be more difficult to find outside of assessments.

8. References

[1] F. K. Goodwin and K. R. Jamison, *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression*. Oxford University Press, 2007.

[2] K. R. Merikangas, H. S. Akiskal, J. Angst, P. E. Greenberg, R. M. Hirschfeld, M. Petukhova, and R. C. Kessler, "Lifetime and 12-month prevalence of bipolar spectrum disorder in the national comorbidity survey replication," *Archives of general psychiatry*, vol. 64, no. 5, pp. 543–552, 2007.

[3] T. Deckersbach, A. A. Nierenberg, M. G. McInnis, S. Salcedo, E. E. Bernstein, D. E. Kemp, R. C. Shelton, S. L. McElroy, L. G. Sylvia, and J. H. Kocsis, "Baseline disability and poor functioning in bipolar disorder predict worse outcomes: Results from the bipolar choice study," *The Journal of clinical psychiatry*, vol. 77, no. 1, pp. 100–108, 2016.

[4] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The Lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.

[5] M. Matthews, G. Doherty, J. Sharry, and C. Fitzpatrick, "Mobile phone mood charting for adolescents," *British Journal of Guidance & Counselling*, vol. 36, no. 2, pp. 113–129, 2008.

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

[7] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Moodscope: building a mood sensor from smartphone usage patterns," in *ACM International Conference on Mobile Systems, Applications, and Services*. ACM, 2013, pp. 389–402.

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

[9] J. Gideon, E. Mower Provost, and M. McInnis, "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016.

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

[11] M. Hamilton, "Hamilton depression scale," *ECDEU Assessment Manual For Psychopharmacology, Revised Edition. Rockville, MD: National Institute of Mental Health*, pp. 179–92, 1976.

[12] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanata, and E. P. Scilingo, "Speech analysis for mood state characterization in bipolar patients," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2012, pp. 2104–2107.

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

[14] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 970–974.

[15] J. Angst, R. Sellaro, and F. Angst, "Long-term outcome and mortality of treated versus untreated bipolar and depressed patients: a preliminary report," *International Journal of Psychiatry in Clinical Practice*, vol. 2, no. 2, pp. 115–119, 1998.

[16] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, T. Gedeon, M. Breakspear, and G. Parker, "A comparative study of different classifiers for detecting depression from spontaneous speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8022–8026.

[17] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," in *Interspeech*, 2013, pp. 2172–2176.

[18] P. Lopez-Otero, L. Dacia-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for depression detection," in *Biometrics and Forensics (IWBF), 2014 International Workshop on*. IEEE, 2014, pp. 1–6.

[19] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 57–63.

[20] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, and M. Graciarena, "The sri avec-2014 evaluation system," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 93–101.

[21] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[23] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition." in *Interspeech*, 2006.

[24] R. Young, J. Biggs, V. Ziegler, and D. Meyer, "A rating scale for mania: reliability, validity and sensitivity." *The British Journal of Psychiatry*, vol. 133, no. 5, pp. 429–435, 1978.

[25] S. O. Sadjadi and J. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[26] S. Tilsen and A. Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 628–639, 2013.

[27] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. Yen, C. C. Tung, and H. H. Liu, "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*, vol. 454, no. 1971, pp. 903–995, 1998.

[28] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 1605–1608.

[29] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[30] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133–147, 1998.

[31] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *2001: A Speaker Odyssey. The Speaker Recognition Workshop*. International Speech Communication Association (ISCA), 2001.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[33] M. Śmieja, "Weighted approach to general entropy function," *IMA Journal of Mathematical Control and Information*, p. dnt044, 2014.

[34] G. Liu and J. H. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 1978–1992, 2014.

[35] L. Ferrer, M. K. Sönmez, and E. Shriberg, "An anticorrelation kernel for improved system combination in speaker verification." in *Odyssey*. Citeseer, 2008, p. 22.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[37] D. J. Miklowitz and M. J. Gitlin, *Clinician's Guide to Bipolar Disorder*. Guilford Publications, 2015.