FEATURES OF VOCAL FREQUENCY CONTOUR AND SPEECH RHYTHM IN BIPOLAR DISORDER

A. Guidi[1, 2], J. Schoentgen[3], G. Bertschy[4], C. Gentili[5], E. P. Scilingo[1, 2], N. Vanello[1, 2]

[1]Dipartimento di Ingegneria dell'Informazione, University of Pisa, Via G. Caruso 16, 56122, Pisa, Italy
[2]Research Center "E. Piaggio", University of Pisa, Largo L. Lazzarino 1, 56122, Pisa, Italy
[3]Dept. of Signals, Images and Acoustics, Faculty of Applied Sciences, Université Libre de Bruxelles, Av. F. Roosevelt 50, 1050, Bruxelles, Belgium
[4]Dept. of Psychiatry, Strasburg University Hospital; Translational Medicine Federation, University of Strasburg; INSERM u1114; 1 place de l'Hôpital, 67000 Strasburg, France
[5]Dept. of General Psychology, University of Padua, Via Venezia 8, 35131, Padua, Italy
andrea.guidi@for.unipi.it, jschoent@ulb.ac.be, gilles.bertschy@chru-strasbourg.fr, c.gentili@unipd.it, e.scilingo@centropiaggio.unipi.it, nicola.vanello@iet.unipi.it

Abstract: **Mental diseases are increasingly common. Among these, bipolar disorders heavily affect patients' lives given the mood swings ranging from mania to depression. Voice has been shown to be an important cue to be investigated in relation with this kind of disease. In fact, several speech-related features have been used to characterize voice in depressed speakers. The goal is to develop a decision support system facilitating diagnosis and possibly predicting mood changes. Recently, efforts were devoted to studies concerning bipolar patients. A spectral analysis of F0-contours extracted from audio recordings of a text read by bipolar patients and healthy control subject is reported. The algorithm is automatic and the obtained features describe parsimoniously speech rhythm and intonation. Bipolar patients were recorded while experiencing different mood states, whereas the control subjects were recorded at different days. Feature trends are detected in bipolar patients across different mood states, while no significant differences are observed in healthy subjects.**
Keywords: **bipolar disorder, mood state, speech contour, speech rhythm, spectral analysis**

1. INTRODUCTION

Mental illnesses have an increasing impact in contemporary society [1]. In particular, the lives of persons suffering from bipolar disorder may be impaired due to periodic, and sometimes extreme, mood swings. Patients may experience oscillations between depression, mania or hypomania, euthymia and a mixed condition [2]. More specifically, depression is a very low mood state characterized by sadness and hopelessness. Mania and hypomania, i.e. a less sever form of mania, are states of hyperarousal that leads to euphoria or irritability, excessive energy, and increased activity. Euthymia is a mood state in which symptoms are mostly absent, and mixed condition is a mood state in which both depressive and manic symptoms are associated. The development of decision support systems may be useful in helping physicians in formulating a diagnosis. With this aim some biomedical studies have been carried out to detect physiological correlates of mood changes [3, 4]. Moreover, several studies have focused on possible relations between voice and mental disease, especially in persons suffering from depression. Indeed, almost the whole central nervous system is involved in voice production [5]. The speech motor system is able to finely control the laryngeal muscles, mucosae, tongue and lips, while the vagus nerve, that is also responsible for the innervation of the motor parasympathetic fibers, activates the pharynx, the soft palate and the laryngeal muscles [5]. In addition, because both the autonomic and somatic nervous systems control the respiratory system, they are expected to affect also the speakers' prosody. As a consequence, speech analysis represents an interesting, non-invasive and economic approach to the study of a speaker's mental state since speech production is modulated by different psychological and/or mental states [6].

Usually three different categories of speech related features can be taken into account in speech analysis: source, vocal tract and prosody-related features [7]. Source features aim at describing the glottal excitation. Vocal tract features report indirectly different shapes of the vocal tract cross-section during articulation. Finally, prosodic features describe the supra-segmental modulation of intonation and intensity of speech. Several studies have been conducted on the latter, because prosody has a role in conveying emotion. Moreover, prosodic feature estimation is fairly robust in noise. Therefore, a system based on a smartphone device, and aiming at analyzing prosodic features in running speech, has been proposed to monitor bipolar patients [8].

Till recently, the study of speech changes in depression was the most often investigated topic [9, 10, 11, 12, 13]. Especially, speaking rate has been found to correlate negatively with the Hamilton Depression Rating Scale score [9]. Five possible descriptors, i.e. decreased speaking intensity, decreased vocal frequency range, slower speech, flat intonation and a lack of linguistic stress, have been highlighted in depressed patients by Hollien [11]. F0 contours were perceptually investigated and demonstrated to contain information about a wide range of prosodic information such as F0 variability, speech rate and pause time [13]. On the contrary, the relationship between mania and speech features has been less explored. More precisely, pressured speech, i.e. the tendency to speak quickly and loudly [14], has been observed to be one of the most common symptoms in children and adolescent affected by mania [15]. Recently, more attention has been paid to the analysis of speech changes in bipolar disorder. To our knowledge, two European Projects investigated speech of patients affected by Bipolar Disorder: namely PSYCHE [3] and MONARCA [4]. As a result, several articles on this topic have been published [16, 17, 18, 19, 20, 21, 22, 23]. Additional studies focused on that topic are [24, 25]. Specifically, intra-subject studies [16, 18] reported significant differences in vocal frequency (F0) variability and average between different mood states. Moreover, the speech intonation contour has been found to be a reliable indicator of mood changes from an euthymic to either a depressed or a manic state in [17, 20]. A system, based on smartphone-sensing, aiming at the recognition of depressive and manic states and the detection of state changes in patients suffering from bipolar disorder was studied in [21, 22, 24]. In these studies, statistics about the phone calls, the verbal interaction of the patients with the other talker, and speech-related features, extracted with the open-source "openSmile" toolbox [26] providing several low-level descriptors, were investigated. A good performance in terms of pattern recognition accuracy was obtained [22]. In [25], a smartphone-based system was proposed to monitor bipolar patients in terms of social rhythms by investigating statistics about conversations, speaking rate and F0 changes. Audio, accelerometer and self-assessment related data were used to classify mood states in bipolar patients in day-to-day phone conversation, achieving an accuracy better than 80% [23].

Despite the relevance of these studies, the analysis of speech changes in mood disorders remains a challenging task. Particularly, the direction of the features changes was not always coherent across subjects [17]. The need of selecting patient specific features and building personalized models was stressed in [22] and [24], respectively. Moreover, in some studies, no significant correlation between F0 variables and depression was found [9, 27]. The improvement of existing models performance could be obtained both by improving subject status characterization, e.g. by evaluating anxiety level [17, 28, 29], and by investigating other features. Novel speech features might in fact capture relevant information about the specific phenomenon under investigation [30]. Furthermore, the exploration of novel speech features [30] that parsimoniously describe the phenomenon of interest in a specific application could improve the classification performance, by avoiding overfitting.

The aim of the present study is to explore a new feature set to characterize speech production in patients affected by bipolar disorder. More precisely, a spectral analysis of the F0 contour is proposed to investigate differences in mood states in patients suffering from bipolar disorder. Specifically, a parsimonious description of the F0-contour is presented. The proposed features are related both to modulation of F0 and to speech rhythm, i.e. speech rate, distance between pauses and pause lengths. Moreover, to better characterize the proposed features, we explore whether and to what extent they are related to rhythm features or whether they represent a complementary source of information. Therefore the rhythmic properties of the audio recordings and a correlational analysis between rhythm features and F0-contour features have been studied. The analysis is carried out in patients and healthy controls. Patients have been recorded at different days, while reading a neutral text. In addition, a study of healthy control speakers is presented. Preliminary results are reported and discussed.

2. METHODS

2.1. Experimental protocol and data

In this study, eleven patients (5 females and 6 males, 40.00 ± 9.02 years) suffering from bipolar disease were enrolled in the Psyche European project [3]. All were able to lead independent and active lives, and they were free from substance use disorder. Seven psychiatric patients were French native speakers, while four of them were Italian native speakers. The experimental protocol, approved by the clinical ethical committee, consisted in the reading of a neutral text during each recording session. Sessions were held at two or three different days. Seven patients out of eleven (patients A-G) (Table 1) were recorded twice each day. A physician labeled the patient's mood status before each recording using clinician-administered rating scales. Four different mood states were identified in this study, namely depressed, euthymic, hypomanic and mixed. A high quality directional microphone was used to record signals at a sampling frequency of 48 KHz and with a resolution of 32 bits (AKG Perception P220 Condenser Microphone, M-Audio Fast-Track).

Eighteen healthy control subjects (9 males and 9 females, 30.00 ± 5.00 years) were enrolled. Healthy control subjects did not report any actual or past psychiatric disorder, and had no history of neurological or major somatic conditions. They were recorded twice at two different days to test for inter-day variability. Typically, the second session took place 7 days after the first one. All healthy subjects were Italian native speakers.

The CMU Arctic Database [31] was used to evaluate the performance of a Voice Activity Detection (VAD) algorithm that is used to classify audio frames. The database provides audio and electroglottographic (EGG) recordings. The EGG signal is a signal that is related to the impedance changes during vocal folds contact. The corpus is formed of about 1100 short sentences, comprising more than 8000 vowels. Audio and EGG recordings were sampled at a rate of 32 kHz with a resolution equal to 32 bit.

2.2. Algorithm

Two different sets of features have been investigated in this study. The first aims at describing the properties of the F0 contour in terms of the shape of its spectrum. The second takes into account rhythm.

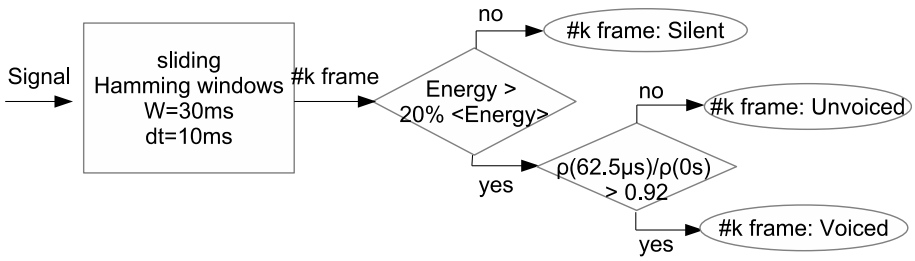

Figure 1: Chart of VAD algorithm, where <·> stands for the average, and ρ is the autocorrelation function.

Spectral shape features. In a first step, VAD, which is inspired by Hess's study [32], is carried out by means of the autocorrelation function (ρ) and speech energy as described in [33]. Audio frames are classified as silent, voiced and unvoiced. Frames characterized by a high relative energy and a high autocorrelation function are labeled as voiced, while low relative energy frames are labeled as silent. Finally, high relative energy but low autocorrelation function frames are labeled as unvoiced (see Figure 1). In a second step, the F0 contour is estimated within each voiced fragment by means of Camacho's Swipe' algorithm [34]. It estimates F0 via the fundamental frequency of the sawtooth waveform whose spectrum best matches the spectrum of the input audio signal. The performance of the F0 estimation algorithm has been assessed via a comparison with other state of the art algorithms in [34, 35, 36], while a validation of the implemented VAD algorithm is proposed in the following. A cubic spline interpolation is used to obtain simulated F0 contours in unvoiced fragments, while F0 within silent pauses is set to 0 Hz (Figure 2). At last, a set of 7 features is obtained from the spectrum of each mean-subtracted F0-contour of the whole utterance. Power spectral density is estimated for each recording via the periodogram. The estimated features are the median frequency (F_{median}), the power amplitude at the median frequency (A_{median}), the maximum peak power amplitude (A_{peak}) and the corresponding frequency (F_{peak}), the ratios between amplitudes and corresponding frequencies as well as the slope according to (Eq. 1-3) (Figure 3).

$$Ratio_{peak} = A_{peak}/F_{peak} \quad (1)$$

$$Ratio_{median} = A_{median}/F_{median} \quad (2)$$

$$\text{Slope} = \frac{A_{peak}-A_{median}}{F_{peak}-F_{median}} \quad (3)$$

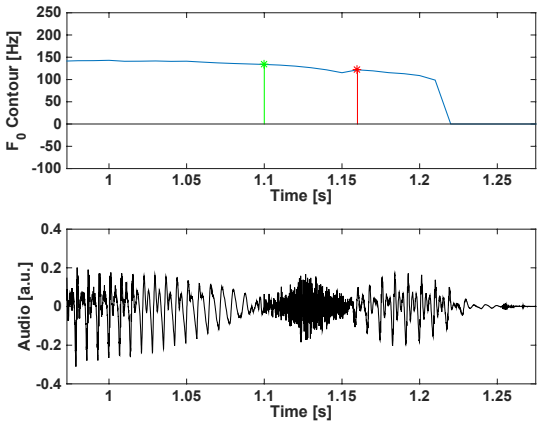

Figure 2: Example of F0 contour. F0 is set to 0 Hz during silent segments, and spline-interpolated during unvoiced segments (marked by green and red lines).

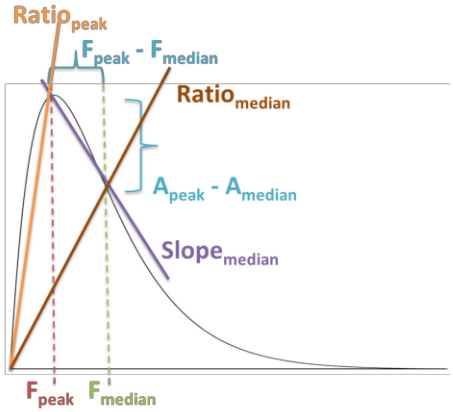

Figure 3: Illustration of the F0-contour power spectral density and of the proposed features.

Rhythm features. The output of the VAD algorithm was used to obtain six rhythm features. According to [37, 38], pauses were categorized into two classes: brief pauses, i.e. the ones having a duration shorter than 200 ms, and medium-long pauses, with a duration longer than 200 ms. In [37] the authors identified also pauses having a duration longer than 1000 ms, but since the authors reported that these are only present in spontaneous speech, they were lumped with the medium pauses (200-1000 ms) in this study. Due to the time resolution of the VAD, pauses shorter than 10 ms cannot be detected. The median duration of the brief pauses ($Dur_{BriefPauses}$), the median duration of the medium-long pauses ($Dur_{LongPauses}$), and the median duration of all the pauses ($Dur_{AllPauses}$) were estimated as well as the median time distance between two consecutive voiced fragments ($Dist_{Voiced}$). The durations of the voiced and unvoiced fragments were combined in the voiced/unvoiced rate (V/UV). Finally, an estimate of the speaking rate was obtained via the number of voiced fragments divided by the duration of the whole utterance, i.e. the voiced fragment rate (VRate).

2.3. VAD algorithm validation

The performance of the VAD method was evaluated by exploiting a database of audio and concurrent electroglottographic (EGG) recordings [31]. Specificity and sensitivity of the VAD were evaluated separately for voiced and silent segments. An overall accuracy was estimated taking also into account unvoiced segments. Segmentation performed via EGG signals was considered as ground truth for the detection of voiced segments, because the EGG signal is directly related to vocal folds contact. Differently, the algorithm proposed by Sohn et al. [39] and based on the likelihood ratio, was used as ground truth segmentation pertaining the detection of silent segments. Finally, the remaining segments were labeled as unvoiced.

2.4. Statistical analysis and Feature Normalization

In this study, intra and inter-subject analyses are carried out. Intra-subject analyses are possible because several recordings of the same subject in different mood states are available. Intra-subject analysis thus enables investigating whether coherent changes can be observed within subjects when switching from one mood state to another. When independent samples are available, i.e. only one observation for each subject, an inter-subject investigation is performed, the goal of which is to study differences in the features values acquired in different mood states. In this case, a possible confounding factor is speaker identity [27, 40]. For this reason, speech features are normalized before inter-subject analysis. Normalization is carried out by dividing speech features obtained in depressive or hypomanic states by those obtained in the euthymic state.

When the same mood state was recorded twice the same day for the same speaker, then the features obtained for the two recordings were averaged for inter-mood comparisons. Friedman's test was used to assess statistically significant differences in paired data corresponding to different mood states in the same patients, and to evaluate the intra-day variability in the patients who carried out the task twice a day. For the latter, recordings at the same acquisition day were not averaged.

A Mann-Whitney U-test was used to assess differences between depression and hypomania for independent samples (i.e., different patients and different mood states). This test was carried out with and without normalization with respect to the same patient's feature data obtained in the euthymic state.

Possible statistical differences due to the language spoken were studied by means of a Mann-Whitney U-test. Patients in the depressed state were compared to test for feature differences due to native language. These tests were applied to normalized data with respect to the euthymic state data. The same comparison was not possible for hypomania due to the small number of subjects. A p-value lower than or equal to 0.05 was considered significant.

A correlational study of the features was performed by means of principal component analysis (PCA) and their representation in the correlation circle. The aim is to highlight whether and to what extent the spectral features share information with the rhythm features. Moreover, the correlation circle enables analyzing which are the features that explain the larger portion of variance in the data. Only the principal components that explain up to 80% of the data variance are taken into account. Both bipolar patients and healthy control subjects were involved in the correlational analysis.

3. RESULTS

Regarding VAD performance, each frame label was compared with the corresponding obtained from the EGG. An overall accuracy equal to 85% was obtained for the classification of the three frame categories, voiced, unvoiced and silent. With regard to voiced segment detection, a specificity equal to 76% and a sensitivity equal to 83% were observed, while a specificity of 82% and a sensitivity of 91% were obtained for unvoiced segment detection.

On average, the reading of the neutral text by bipolar patients and control subjects lasted about 4 minutes. Each bipolar patient but one, i.e. patient E, reported a euthymic state in one of the recording days (Table 1). Table 1 enables selecting the subjects that can be used for both paired and independent data tests. Some features share the same behavior for all subjects. In particular, F_{peak} was always lower than F_{median}, thus resulting in negative Slope values and in a $Ratio_{peak}$ that was always higher than $Ratio_{median}$.

A paired test was applied to data of patients A, B, C and G to investigate possible statistically significant differences between hypomanic and euthymic states, while data of patients B, D, F, H, I, L and M were used to study statistically differences between depressed and euthymic states.

Mann-Whitney U-test for independent data was applied to compare features obtained from patients in a depressed state, i.e. D, F, H, I, L, M, with features of patients in a hypomanic state, i.e. A, B, C, G.

Inter-language tests for independent data were applied to compare patients in depressed state. Specifically, patients H, I, L, M and B, D F were involved in an inter-language, but intra-depressed state study.

Table 1: Patient mood labels.

subj.	Label day1	Label day2	Label day3
A	Hypomania	Euthymia	
B	Hypomania	Euthymia	Depression
C	Hypomania	Euthymia	
D	Depression	Euthymia	
E	Depression	Hypomania	
F	Depression	Euthymia	
G	Hypomania	Euthymia	
H	Depression	Euthymia	
I	Depression	Euthymia	
L	Depression	Euthymia	
M	Mixed	Depression	Euthymia

3.1. Results for spectral shape features

Analysis of paired data (Table 2) shows statistically significant differences between hypomania and euthymia states for A_{peak}, F_{peak}, $Ratio_{peak}$ and Slope (patients A, B, C and G). For all subjects (but one), F_{peak} (Figure 4) as well as Slope (Figure 5) were lower in the hypomanic state, while for all the subjects A_{peak} (Figure 4) and $Ratio_{peak}$ (Figure 4) were higher in the hypomanic state.

Table 2: Bipolar patients – spectral shape features: p-values < 0.05 are in bold.

	Test	F_{median}	A_{median}	F_{peak}	A_{peak}	Slope	$Ratio_{peak}$	$Ratio_{median}$
Hyp. vs. Eut.	Paired	3.17E-01	3.17E-01	**4.55E-02**	**4.55E-02**	**4.55E-02**	**4.55E-02**	3.17E-01
Dep. vs. Eut.	Paired	**8.15E-03**	7.05E-01	7.05E-01	**8.15E-03**	**8.15E-03**	7.05E-01	2.57E-01
Dep. vs. Hyp.	UnPaired	**9.52E-03**	1.71E-01	3.52E-01	4.76E-01	6.67E-02	4.76E-01	2.57E-01

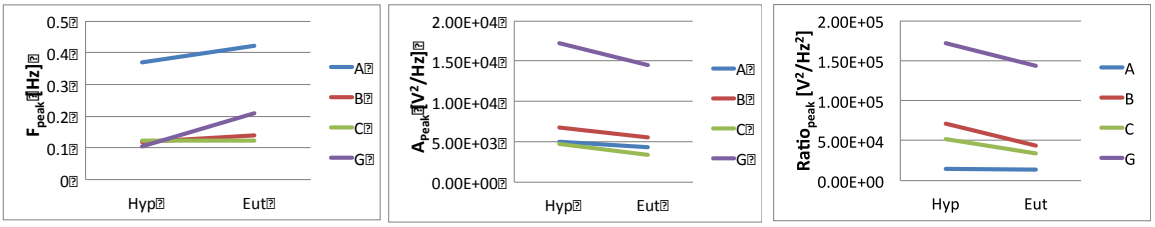

Figure 4: F_{peak} (left), A_{peak} (center), and $Ratio_{peak}$ (right) in patients switching from hypomania to euthymia.

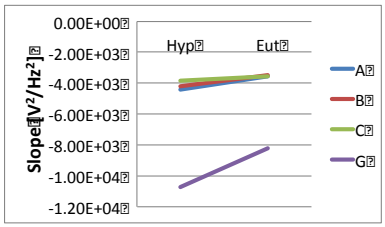

Figure 5: Slope in patients switching from hypomania to euthymia.

Analysis of paired data (patients B, D, F, H, I, L and M) show that differences between depression and euthymia were statistically significant for F_{median}, A_{peak} and Slope. For all subjects, F_{median} (Figure 6) and Slope (Figure 6) were lower in the depressed state, while A_{peak} (Figure 6) was higher.

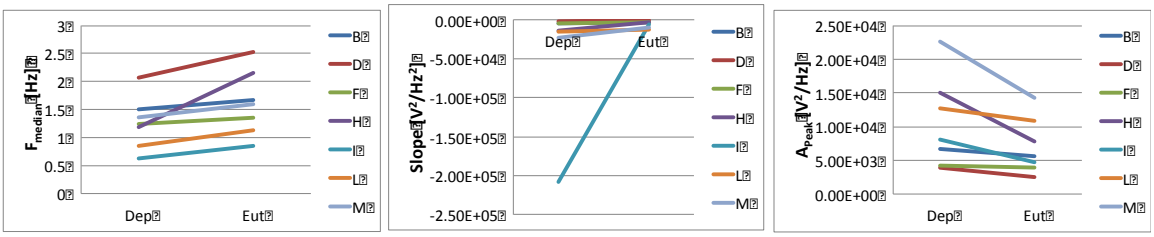

Figure 6: F_{median} (left), Slope (center) and A_{peak} (right) in patients switching from depression to euthymia.

Comparisons carried out via the Mann-Whitney U-test on unpaired normalized data between depression and hypomania (Table 2) showed statistically significant differences for F_{median} (Figure 7). F_{median} was lower in the depressed state with respect to hypomania. Without normalization, the features did not show any statistically significant differences. In the box-plots reported in this study, median values are marked by a change in color.

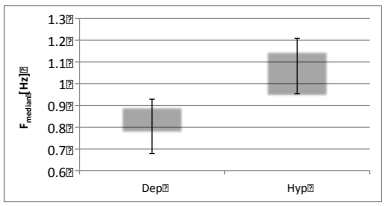

Figure 7: Boxplot of normalized F_{median} in patients switching from depression to hypomania.

3.2. Results for rhythm features

Statistical pairwise analysis of the rhythm features (Table 3) showed statistically significant differences for $Dur_{LongPauses}$, $Dur_{AllPauses}$ as well as V/UV, when comparing hypomania and euthymia, and in $Dur_{LongPauses}$, $Dist_{Voiced}$ as well as VRate, when comparing depression and euthymia. $Dur_{LongPauses}$ is significantly lower (Figure 8) in the hypomanic state, and significantly higher (Figure 9) in the depressed state with respect to euthymia. Similarly, $Dur_{AllPauses}$ is lower (Figure 8) in the hypomanic state with respect to euthymia. Moreover, the V/UV ratio (Figure 8) is significantly higher in hypomania. $Dist_{Voiced}$ (Figure 9) is higher in the depressed state, while the opposite is observed for VRate (Figure 9).

Table 3: Bipolar patients – rhythm features: p-values < 0.05 are in bold.

	Test	$Dur_{BriefPauses}$	$Dur_{LongPauses}$	$Dur_{AllPauses}$	V/UV	$Dist_{Voiced}$	VRate
Hyp. vs. Eut.	Paired	3.17E-01	**4.55E-02**	**4.55E-02**	**4.55E-02**	3.17E-01	3.17E-01
Dep. vs. Eut.	Paired	2.56E-01	**8.15E-03**	5.88E-02	2.57E-01	**8.15E-03**	**8.15E-03**
Dep. vs. Hyp.	UnPaired	6.10E-01	**1.90E-02**	**9.50E-03**	2.57E-01	**3.81E-02**	**1.90E-02**

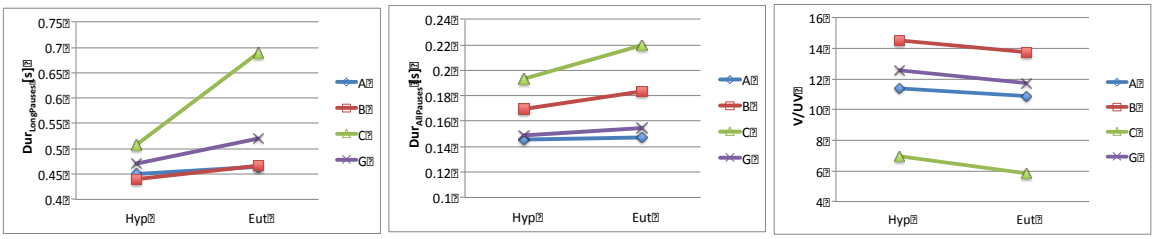

Figure 8: $Dur_{LongPauses}$ (left), $Dur_{AllPauses}$ (center) and V/UV (right) in patients switching from hypomania to euthymia.

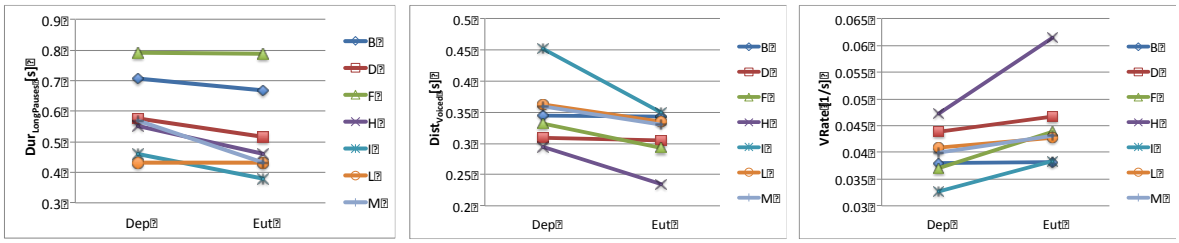

Figure 9: $Dur_{LongPauses}$ (left), $Dist_{Voiced}$ (center) and VRate (right) in patients switching from depression to euthymia.

Analyses of unpaired normalized data of depression and hypomania (Table 3) show statistically significant differences for $Dur_{LongPauses}$ (Figure 10), $Dur_{AllPauses}$ (Figure 10), $Dist_{Voiced}$ (Figure 11), and VRate (Figure 11). $Dur_{longPauses}$, $Dur_{AllPauses}$ and $Dist_{Voiced}$ were higher in depression than hypomania, while VRate was higher in hypomania than depression. No statistically significant difference was found without data normalization.

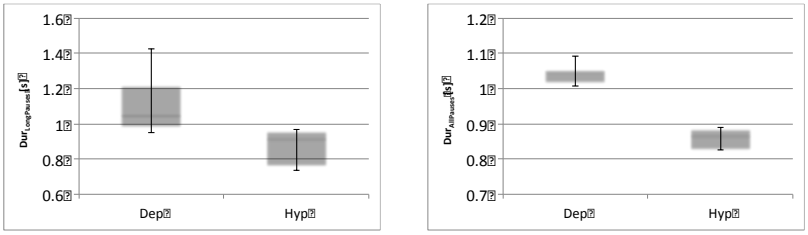

Figure 10: Boxplot of normalized $Dur_{LongPauses}$ (left) and $Dur_{AllPauses}$ (right) in patients switching from depression to hypomania.

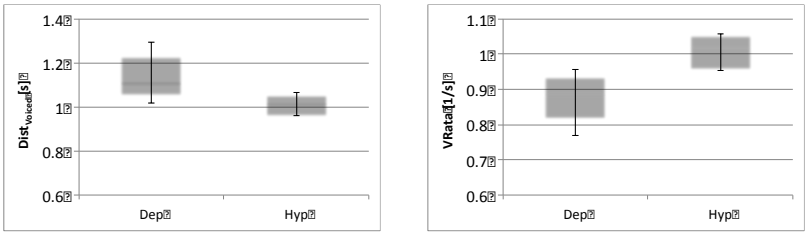

Figure 11: Boxplot of normalized $Dist_{Voiced}$ (left) and VRate (right) in patients switching from depression to hypomania.

3.3. Test for intra-day variability

Seven patients, recorded twice at each acquisition day, were taken into account comparing same-label data. Friedman's test for paired data did not show any statistically significant differences between the two audio recordings carried out the same day (data not shown).

3.4. Test for inter-language variability

Possible differences in feature trends, depending on language, were evaluated by applying the Mann-Whitney U-test (data not shown). The tests did not report any statistically significant differences when applied to normalized data .

3.5. Results for healthy control subjects

Analysis of the data of the healthy control subjects did not return statistically significant differences between features obtained from audio samples recorded at two different days. The p-values are reported in Tables 4 - 5, while the F_{median} trends in healthy control subjects are displayed in Figure 12.

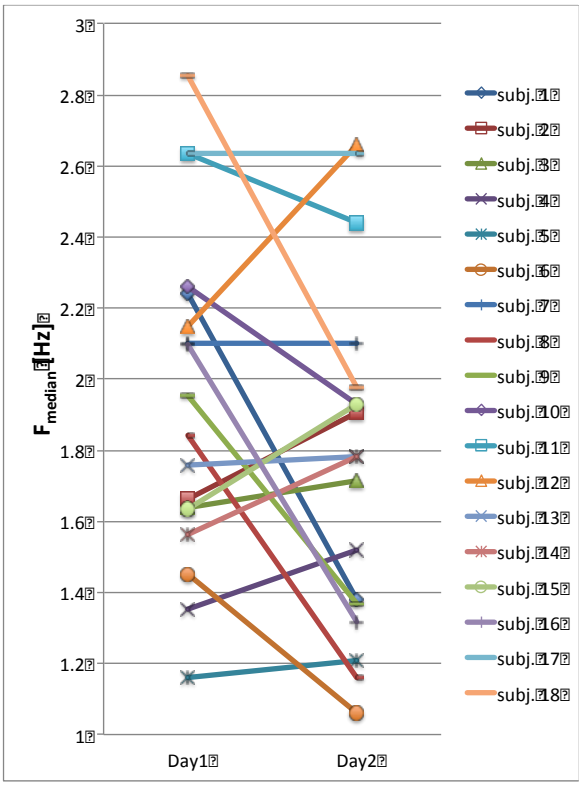

Figure 12: F_{median} trends in healthy control subjects.

Table 4: Healthy control subjects – spectral shape features: p-values.

F_{median}	A_{median}	F_{peak}	A_{peak}	Slope	$Ratio_{peak}$	$Ratio_{median}$
3.17E-01	3.46E-01	3.46E-01	6.37E-01	6.37E-01	6.37E-01	3.46E-01

Table 5: Healthy control subjects – rhythm features: p-values.

$Dur_{BriefPauses}$	$Dur_{LongPauses}$	$Dur_{AllPauses}$	V/UV	$Dist_{Voiced}$	VRate
3.46E-01	4.67E-01	6.37E-01	1.00E+00	1.00E+00	3.46E-01

3.6. Correlational study

An analysis of the principal components (PCs) shows that the first component (PC1) is able to explain 41.80% of the total variability, PC2 explains 18.18%, PC3 13.05% and PC4 the 11.13%. Therefore, a total variability equal to 84.16% is explained by the first four PCs. Two circles of the correlations, in which the original variables are located as a function of the correlation between variables and principal components (i.e. loadings), are shown in Figure 13. In Figure 13 the principal component space defined by PC1 and PC2 and by PC3 and PC4 are shown. The analysis reveals that $Ratio_{peak}$ and A_{peak} are at the same time very close to each other and very close to the unit circle which means that those variables share the same behavior and are likely to be correlated. The Pearson's correlation (ρ) coefficient is equal to 0.93. The same property holds for $Ratio_{median}$ and A_{median} (ρ =0.99) and Slope (correlates with A_{median} ρ = –0.5 and with $Ratio_{median}$ ρ = –0.59). VRate and $Dist_{Voiced}$ show similar loadings with respect to PC1 and PC2, but with opposite signs (ρ =-0.91). Similarly, A_{peak} and VRate show similar loadings with respect to PC1 but opposite with respect to PC2 (ρ =0.48), while A_{peak} and $Dist_{Voiced}$ show opposite loadings with respect to PC1 but similar with respect to PC2 (ρ =-0.44). Again, Slope shows to be positively correlated with VRate (ρ =0.34) and negatively with $Dist_{Voiced}$ (ρ =-0.36). Only VRate and F_{median} show similar loadings in the PC1-PC2 plane (ρ =0.56) when exploring correlations between spectral and rhythm features. In the PC3-PC4 plane (Figure 13), $Dur_{LongPauses}$

and F_{peak} reveal similar loadings but with opposite signs. These two variables are mainly related to PC3. PC4 seems to be mainly related to the V/UV ratio and only weakly related to $Dur_{LongPauses}$ and F_{peak}.

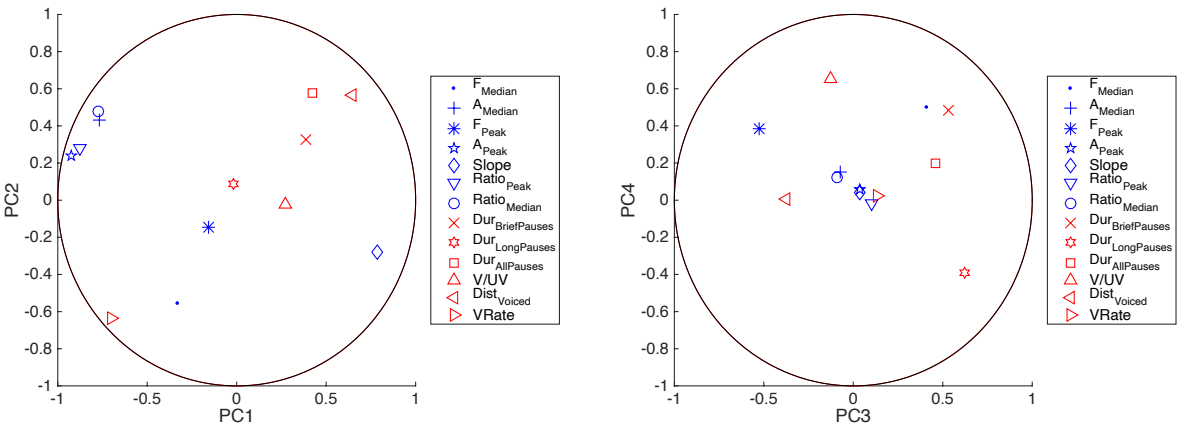

Figure 13: Correlation circle – PC1 vs. PC2 (left) and PC4 vs. PC3 (right). Spectral shape features are reported in blue, while rhythm features are in red.

4. DISCUSSION

In this study an automatized spectral analysis of F0-contours is carried out. Conventionally, the F0-contour is studied in the time domain. We demonstrated that an analysis in the frequency domain may provide a compact description of the F0-related prosodic information. Because F0 was set to zero in silent fragments, spectral features summarize the contribution of rhythm as well as intonation. In particular, results depend not only on syllabic rhythm (4Hz typically), but also on pauses between words or sentences. This has been confirmed by the PCA and correlation analysis we performed taking into account rhythm features. Specifically, a significant correlation was found among speech rate features such as VRate and $Dist_{Voices}$, and spectral shape features such as F_{median}, A_{peak} and Slope. Moreover, the PCA analysis revealed that the two families of variables, related to the spectral description of the F0-contour and speech rhythm respectively, carry complementary information. In fact, they are differently related to the PCs that together describe a large percentage of the variance.

The performance of the VAD algorithm was also evaluated. Although the performance of this processing step can be improved, the proposed algorithm shows interesting properties in terms of ease of implementation. Such a property might enable a battery-saving implementation on mobile platforms, such as smartphone devices.

Statistical analyses were carried out on speech features of bipolar patients experiencing different mood states, and on control subjects recorded at different days. Statistically significant differences were found between features across different mood states. Some features showed a good specificity, insofar they were similar for control subjects recorded at different days and for patients recorded in the same state at the same day.

With regard to the paired data analysis of the spectral shape features, a comparison of euthymia and depression showed that A_{peak} increases and F_{median} and Slope decrease in depression (Figure 6). Since F_{peak} is lower than F_{median} this indicates a larger contribution of the low frequency spectral components in depression. Preliminary results from independent samples show a decrease of F_{median} in depressed with respect to hypomanic patients (Figure 7). Given the correlation observed between F_{median} and speaking rate, VRate in particular (Figure 11), these results are in good agreement with the literature. Some studies indeed report a decrease of speaking rate in depressed patients [9, 11], while an increase was observed in patients affected by mania [15]. Since F_{median} takes also into account vocal frequency changes, these results might also reflect a concomitant decrease of F0. Interestingly, F0 was often observed to decrease during depressive episodes [27] and to increase in patients experiencing hypomania [18]. The paired analysis of the hypomanic and euthymic states revealed a significant increase of A_{peak} (Figure 4) and a decrease of F_{peak} (Figure 4) as well as of Slope (Figure 5) in hypomania, while no coherent change of F_{median} was observed. Since A_{peak} is positively correlated with the VRate and negatively with the $Dist_{Voiced}$, these results indicate a specific prosodic dynamic that is compatible with a pressured speech.

These results may suggest that A_{peak} and Slope may be used as a "no symptoms marker" to discriminate euthymia from both depression and hypomania. Again, these findings may indicate that F_{median} may be used to discriminate

depression from both euthymia and hypomania. Although these mood states may appear easily differentiable, such a result may not be trivial since both depression and mania can share symptoms, e.g. psychomotor agitation [41], that might confound their speech-based recognition.

With regard to statistical tests on independent samples, it is important to stress that statistically significant results were obtained only after normalizing the feature values. We believe that the normalization step is necessary to mitigate the effect of speaker identity. Confounding factors in inter-subject comparisons could be related to language but also to speaking style. Since we were interested in highlighting possible differences in the speech features due to mood change, we decided to normalize with respect to the euthymic state. This normalization was performed assuming that euthymia is the emotional point of reference since it lacks relevant symptoms. The analysis of inter-linguistic differences revealed no statistically significant differences between Italian and French speakers when the features were normalized with reference to euthymia. The population size here involved allowed to carry out this test with depressed subjects.

Notwithstanding the small number of patients who have been analyzed, we perceive the results to be relevant because common feature trends have been detected in patients across mood states. Due to the sample size, it was not possible to perform any statistical tests on paired data involving depression and hypomania. Also, we believe that the results are relevant because they report a novel feature set that parsimoniously describes the changes in bipolar patients' speech. The proposed features are related to changes in the modulation of the vocal frequency and to speech rhythm.

5. Conclusion

This study has investigated a spectral analysis of F0-contours to characterize mood changes in bipolar patients. The features report the modulation of the vocal frequency as well as speech rhythm. The small number of enrolled patients does not enable generalizing the results, but observed trends suggest that the proposed features may warrant further study. Statistically significant differences were discovered between features describing a neutral text read by bipolar patients in different mood states. Common feature trends were observed in enrolled patients across mood states. But, no statistically significant differences were found in healthy control subjects from one day to the next or between patients in the same mood state at different times of the day or when speaking different languages.

The proposed approach is automatic and the features enable a parsimonious description of prosodic changes between mood states that could be integrated in a decision support system to help clinicians in the difficult task of making diagnosis and tailoring treatments.

Acknowledgements

This research is partially supported by the EU Commission under contract ICT-247777 Psyche.

References

[1] 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, 367(9524), 2006, pp. 1747-1757.

[2] K.R. Merikangas, R. Jin, J.P. He, R.C. Kessler, S. Lee, N.A. Sampson, M.C. Viana, L.H. Andrade, C. Hu, E.G Karam, and M. Ladea. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Archives of general psychiatry, 68(3), 2001, pp. 241-251.

[3] G. Valenza, C. Gentili, A. Lanatà, and E.P. Scilingo. "Mood recognition in bipolar patients through the PSYCHE platform: preliminary evaluations and perspectives." Artificial intelligence in medicine 57, 1, 2013, pp. 49-58.

[4] O. Mayora, B. Arnrich, J. Bardram, C. Drager, A. Finke, M. Frost, S. Giordano et al. "Personal health systems for bipolar disorder Anecdotes, challenges and lessons learnt from MONARCA project." In Pervasive computing technologies for healthcare (PervasiveHealth), 2013 7th international conference on, pp. 424-429. IEEE, 2013.

[5] J. Kreiman and D. Sidtis, Foundations of voice studies: An interdisciplinary approach to voice production and perception. John Wiley & Sons, 2011.

[6] C. S. Hopkins, R. J. Ratley, D. S. Benincasa, and J. J. Grieco, "Evaluation of voice stress analysis technology," in System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on. IEEE, 2005, pp. 20b–20b.

[7] S.G. Koolagudi, K.S. and Rao. Emotion recognition from speech: a review. International journal of speech technology, 15(2), 2012, pp. 99-117.

[8] A. Guidi, S. Salvi, M. Ottaviano, C. Gentili, G. Bertschy, D. de Rossi, E.P. Scilingo, and N. Vanello. Smartphone Application for the Analysis of Prosodic Features in Running Speech with a Focus on Bipolar Disorders: System Performance Evaluation and Case Study. Sensors, 15(11), 2015, pp. 28070-28087.
[9] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder. Voice acoustical measurement of the severity of major depression. Brain and cognition, 56, 1, 2004, pp. 30-35.
[10] K.E.B. Ooi, M. Lech, and N.B. Allen. Multichannel weighted speech classification system for prediction of major depression in adolescents. Biomedical Engineering, IEEE Transactions on, 60, 2, 2013, pp. 497-506.
[11] H. Hollien. Vocal indicators of psychological stress. Annals of the New York Academy of Sciences, 347(1), 1980, pp. 47-72.
[12] A. Nilsonne, J. and Sundberg. Differences in ability of musicians and nonmusicians to judge emotional state from the fundamental frequency of voice samples. Music Perception: An Interdisciplinary Journal, 2(4), 1985, pp. 507-516.
[13] Nilsonne, Å., Sundberg, J., Ternström, S., and Askenfelt, A. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. The Journal of the Acoustical Society of America, 83(2), 1988, pp. 716-728.
[14] Trzepacz, Paula T., and Robert W. Baker. The psychiatric mental status examination. Oxford University Press, 1993.
[15] Kowatch, R. A., Youngstrom, E. A., Danielyan, A., and Findling, R. L. Review and meta-analysis of the phenomenology and clinical characteristics of mania in children and adolescents. Bipolar disorders, 7(6), 2005, pp. 483-496.
[16] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanatà, and E.P. Scilingo. Speech analysis for mood state characterization in bipolar patients. In Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, pp. 2104-2107, IEEE.
[17] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, and E.P. Scilingo. "Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients." Biomedical Signal Processing and Control, 17, 2015, pp. 29-37.
[18] A. Guidi, E.P. Scilingo, C. Gentili, G. Bertschy, L. Landini, and N. Vanello. Analysis of running speech for the characterization of mood state in bipolar patients. In 2015 AEIT International Annual Conference (AEIT), 2015, pp. 1-6, IEEE.
[19] Guidi, A., Schoentgen, J., Bertschy, G., Gentili, C., Landini, L., Scilingo, E.P. and Vanello, N., 2015, August. Voice quality in patients suffering from bipolar disease. In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, pp. 6106-6109, IEEE.
[20] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, and E.P. Scilingo. An Automatic Method for the Analysis of Pitch Profile in Bipolar Patients. In Proc. 8th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications, 2013, pp. 231- 234, Firenze University Press.
[21] Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Oehler, S., Tröster, G., ... and Lukowicz, P. Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE Journal of Biomedical and Health Informatics, 19(1), 2015, pp. 140-148.
[22] Muaremi, A., Gravenhorst, F., Grünerbl, A., Arnrich, B., and Tröster, G. 2014, May. Assessing bipolar episodes using speech cues derived from phone calls. In International Symposium on Pervasive Computing Paradigms for Mental Health, pp. 103-114, Springer International Publishing.
[23] Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., and Morales, E. F. (2016). Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive and Mobile Computing.
[24] Karam, Z. N., Provost, E. M., Singh, S., Montgomery, J., Archer, C., Harrington, G., and Mcinnis, M. G. 2014, May, Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4858-4862, IEEE.
[25] Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., and Choudhury, T. Automatic detection of social rhythms in bipolar disorder. Journal of the American Medical Informatics Association, 23(3), 2016, pp. 538-543.
[26] Eyben, F., Wöllmer, M., and Schuller, B. ,2010, October, Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, pp. 1459-1462, ACM.
[27] 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, 71, 2015, pp. 10-49.

[28] E. Moore, M.A. Clements, J.W. Peifer, L. and Weisser. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. Biomedical Engineering, IEEE Transactions on, 55(1), 2008, pp. 96-107.
[29] E. Gilboa-Schechtman, L. Galili, Y. Sahar, and O. Amir. Being “in” or “out” of the game: subjective and acoustic reactions to exclusion and popularity in social anxiety. Biased Cognitions & Social Anxiety: Building a Global Framework for Integrating Cognitive, Behavioral, and Neural Processes, 2015.
[30] Muthusamy, H., Polat, K., and Yaacob, S. Particle Swarm Optimization Based Feature Enhancement and Feature Selection for Improved Emotion Recognition in Speech and Glottal Signals. PloS one, 10(3), 2015, pp. e0120344.
[31] Kominek, J., and Black, A. W. (2004). The CMU Arctic speech databases. In Fifth ISCA Workshop on Speech Synthesis.
[32] Hess, W. (1975, September). Time-Domain, Digital Segmentation Of Connected Natural Speech. In IJCAI (pp. 491-498).
[33] J.L Blanco, J. Schoentgen, and C. Manfredi. Vocal tract settings in speakers with obstructive sleep apnea syndrome. In Proc. 8th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications, 2013, pp. 211-214, Firenze University Press.
[34] A. Camacho, and J.G. Harris. "A sawtooth waveform inspired pitch estimator for speech and music." The Journal of the Acoustical Society of America 124, 3, 2008, pp. 1638-1652.
[35] K. Evanini, and C. Lai. The importance of optimal parameter setting for pitch extraction. The Journal of the Acoustical Society of America, 128(4), 2010, pp. 2291-2291.
[36] N. Vanello, N. Martini, M., Milanesi, H. Keiser, M. Calisti, L. Bocchi, C. Manfredi, and L. Landini. Evaluation of a pitch estimation algorithm for speech emotion recognition. In Proc. 6th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications, 2009, pp. 29-32, Firenze University Press.
[37] Campione, E., and Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In Speech prosody 2002, international conference.
[38] J. Fletcher. Some micro and macro effects of tempo change on timing in French. Linguistics, 25(5), 1987, pp. 951-968.
[39] Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. IEEE signal processing letters, 6(1), 1-3.
[40] N. Cummins, J. Epps, M. Breakspear, and R. Goecke. An investigation of depressed speech detection: Features and normalization. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2011, pp. 2997-3000.
[41] Goldberg, J. F., Perlis, R. H., Bowden, C. L., Thase, M. E., Miklowitz, D. J., Marangell, L. B., ... and Sachs, G. S. Manic symptoms during depressive episodes in 1,380 patients with bipolar disorder: findings from the STEP-BD. American Journal of Psychiatry, 166(2), 2009, pp. 173-181.