Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients

A. Guidi[a,*], N. Vanello[a], G. Bertschy[b], C. Gentili[c], L. Landini[a], E. P. Scilingo[a]

[a] A. Guidi, N. Vanello, L. Landini, E. P. Scilingo are with University of Pisa, Dipartimento di Ingegneria dell'Informazione, Pisa, Italy - Via G. Caruso 16 - 56122 - Pisa and with Research Center "E.Piaggio", University of Pisa, Pisa, Italy -Largo Lucio Lazzarino 1 - 56122 - Pisa, Italy.
E-mail: andrea.guidi@for.unipi.it, nicola.vanello@iet.unipi.it, luigi.landini@iet.unipi.it, e.scilingo@centropiaggio.unipi.it
[b] G. Bertschy is with University Hospital and University of Strasbourg, INSERM u666, Department of Psychiatry and Mental Health, Strasbourg, France - 4 rue Blaise Pascal – CS 90032 – F-67081 Strasbourg cedex - 2010.
E-mail: gilles.bertschy@chru-strasbourg.fr
[c] C. Gentili is with University of Pisa, Department of Surgical, Medical, Molecular Pathology and Critical Care, Pisa, Italy - Via Savi 10 - 56126 – Pisa, Italy. E-mail: claudio.gentili@med.unipi.it

Abstract

Bipolar disorders are characterized by a mood swing, ranging from mania to depression. A system that could monitor and eventually predict these changes would be useful to improve therapy and avoid dangerous events. Speech might convey relevant information about subjects' mood and there is a growing interest to study its changes in presence of mood disorders. In this work we present an automatic method to characterize fundamental frequency (F0) dynamics in voiced part of syllables. The method performs a segmentation of voiced sounds from running speech samples and estimates two categories of features. The first category is borrowed from Taylor's Tilt intonational model. However, the meaning of the proposed features is different from the meaning of Taylor's ones since the former are estimated from all voiced segments without performing any analysis of intonation. A second category of features takes into account the speed of change of F0. In this work, the proposed features are first estimated from an emotional speech database. Then, an analysis on speech samples acquired from eleven psychiatric patients experiencing different mood states, and eighteen healthy control subjects is introduced. Subjects had to perform a text reading task and a picture commenting task. The results of the analysis on the emotional speech database indicate that the proposed features can discriminate between high and low arousal emotions. This was verified both at single subject and group level. An intra-subject analysis was performed on bipolar patients and it highlighted significant changes of the features with different mood states, although this was not observed for all the subjects. The directions of the changes estimated for different patients experiencing the same mood swing, were not coherent and were task-dependent. Interestingly, a single-subject analysis performed on healthy controls and on bipolar patients recorded twice with the same mood label, resulted in a very small number of significant differences. In particular a very good specificity was highlighted for the Taylor-inspired features and for a subset of the second category of features, thus strengthening the significance of the results obtained with patients. Even if the number of enrolled patients is small, this work suggests that the proposed features might give a relevant contribution to the demanding research field of speech-based mood classifiers. Moreover, the results here presented indicate that a model of speech changes in bipolar patients might be subject-specific and that a richer characterization of subject status could be necessary to explain the observed variability.

Keywords: F0 contour analysis; bipolar disorders; emotional speech; voiced/unvoiced segmentation

1. Introduction

Bipolar disorder is a pathology characterized by cyclic variations of mood. Subjects can experience very different states ranging from hypomania to depression, passing through euthymia. Physicians would strongly benefit from a tool that could support them in formulating diagnoses and monitor patients' status between two successive visits when they are not hospitalized. Such a support system could alert physicians if patient status gets worst, optimize therapy and reassure patients. Several studies are concerned with the analysis of biomedical signals to detect physiological correlates of mood changes [1,2,3]. In particular, the analysis of speech signals could be used to obtain relevant information about patient mood state. In fact, speech-derived characteristics have been shown to vary in patients affected by psychiatric disorders with respect to healthy subjects. Currently, different categories of features are studied including glottal, prosodic, spectral and energy-related features. Each category of features can be tested on its own or in a combined approach. Prosodic and spectral features have been found to vary in patients suffering from depression with respect to healthy subjects [4,5,6,7,8,9,10]. Prosodic, glottal, cepstral, spectral and Teager Energy Operator (TEO) related features were used in detecting depression in adolescents [8] by analysing speeches during family interaction. Although good classification results were found combining the different categories of features, TEO was found to carry the most relevant information. In a related work, Ooi et al. [9] reported that prosodic and glottal features were effective in predicting the evolution of depression in adolescents over a time range of two years. Moore et al. [7] highlighted the relevance of glottal tract features for depression classification and showed that glottal features can improve depressed speech classification more than vocal tract features when used in combination with prosodic features.
The analysis of speech features has also been proposed with the aim of identifying different emotions in speech [11,12]. Scherer sustained that speaker's emotional arousal is associated with physiological changes in respiration, phonation and articulation [13].

* Corresponding author. Tel.: +39 050 2217462; fax: +39 050 2217050; e-mail: andrea.guidi@for.unipi.it

Such changes are responsible of the production of the emotion-specific patterns in acoustic parameters. Generally prosodic features as fundamental frequency, duration, intensity, and less frequently voice quality features as harmonics-to-noise ratio, jitter and shimmer are investigated in the field of emotion recognition [14]. The great availability of low-level descriptors and functionals has encouraged the use of a great number of features (e.g. brute-force extraction) carrying to the implementation of features selection methods [14,15].
At the same time there is an interest in developing models that can improve the description of the dynamics of speech features. Recently, the relevance of shape, slope and range of F0 contour in emotional speech perception, synthesis and recognition has been described [11,16,17,18]. Moreover, local prosodic features that are related to the temporal dynamics description of prosody have been found to improve the information carried by global, static prosodic features [18].
Rising and falling segments of the stylized fundamental frequency were taken into account in [19], where it was highlighted how F0 slope tends to be steeper in higher aroused emotions. The phenomenon of F0 declination across an utterance was studied in emotional speeches [20]. Moreover, it was found that the F0 slope in the last syllable can convey different moods [21].
In this work we want to investigate whether a detailed description of F0 dynamics could be used to discriminate among different emotions in speech, and distinguish different mood states in bipolar patients. To achieve this goal we describe an automatic method for the analysis of F0 contour. The proposed method performs an automatic segmentation of running speech and the detection of voiced parts of syllables. A descriptive statistics of the F0 profile within each voiced segment is suggested. In particular, two categories of features are proposed. The first is borrowed from Taylor's Tilt Intonational Model [22] and it describes geometrically the voiced segments. Unlike the Taylor's model, the features here proposed are investigated in every voiced segment and not in only intonational events, as Taylor postulated in his model. The second category of features is related to the speed of F0 variations and estimates the steepness of both rising and falling F0 trend in each voiced segment. The results obtained from an emotional speech database are shown and preliminary results on bipolar patients, recorded in different mood states, are introduced and discussed.

2. Material and methods

2.1 Features extraction algorithm

The proposed approach is a three-step procedure, consisting in a voiced segments detection step, a speech fundamental frequency (F0) estimation step and a final feature estimation step. In the first step, speech signal is analysed to detect voiced part of syllables by using information about zero crossing rate [23] and signal intensity [24]. More details can be found in Fig. 1.

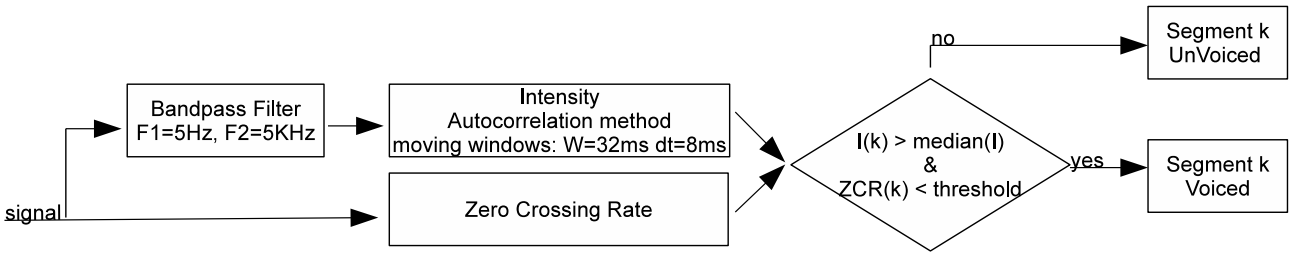

Figure 1: Flowchart of the voiced segment detection step. Only the segments having high intensity and low zero crossing rate are considered voiced.

In a second step a procedure based on the Camacho's swipe' algorithm [25], is used to estimate the F0 contour. As can be seen in Fig. 2, for each voiced segment, the fundamental frequency is estimated using a moving window approach [26].

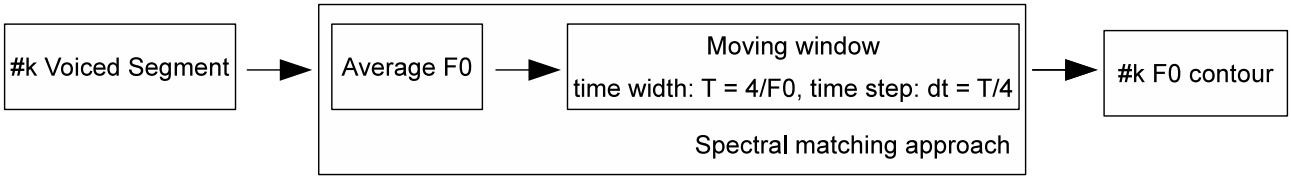

Figure 2: Flowchart of the F0 contour estimation step. The spectral matching approach was performed by using the Camacho's Swipe' algorithm.

In a third step the final features, that allow describing specific characteristics of the F0 profile within each voiced segment, are estimated. In particular some of the extracted features are borrowed from Taylor's Tilt Model [22] and are related to "relative sizes of the amplitude and durations of rises and falls for an event". Within each voiced segment an eventual local maximum is detected and the features are estimated as in (1-3):

$$Amplitude^* = \left(|A_{rise}| - |A_{fall}|\right)/\left(|A_{rise}| + |A_{fall}|\right) \quad (1)$$
$$Duration^* = \left(|D_{rise}| - |D_{fall}|\right)/\left(|D_{rise}| + |D_{fall}|\right) \quad (2)$$
$$Tilt^* = (Amplitude^* + Duration^*)/2 \quad (3)$$

where A_{rise} and A_{fall} are the F0 changes during the rising and falling section within a segment respectively, D_{rise} and D_{fall} are the duration of the rising and falling sections. The features we are estimating are different by those proposed by Taylor, even if functionally equivalent. This is due to the voiced segment and the identification processes utilized. In section 4 this issue will be discussed further. Amplitude* (ampl*) feature is an index of the difference between the F0 amplitude excursions during rising and

falling trend. Duration* (dur*) instead takes into account the time intervals in which the two trends, rising and falling, happen. Finally tilt* is the mean value of the two previous features.
The previously considered features can describe the shape of F0 contour in voiced segments, while they are insensitive to the temporal scale of the phenomena. Thus, a second category of features that takes into account the speed of F0 change has been considered. In particular, the steepness of the F0 contour during both rising (PosSlope) (4) and falling (AbsNegSlope) (5) phase of F0 change is estimated.

$$PosSlope = |A_{rise}|/|D_{rise}| \quad (4)$$
$$AbsNegSlope = |A_{fall}|/|D_{fall}| \quad (5)$$

Finally, other two features are estimated according to (6,7):

$$SumDer = Slope_{rise} + Slope_{fall} \quad (6)$$
$$GlobalSlope = \left(|A_{rise}| - |A_{fall}|\right)/\left(|D_{rise}| + |D_{fall}|\right) \quad (7)$$

SumDer (6) is estimated as the sum between the F0 slope during the rising variation and the F0 slope during the falling variation. GlobalSlope (7) instead is defined as the F0 slope between the first and the final F0 values in each voiced segment.

2.2 Experimental protocol

2.2.1 Segmentation performances test

The performances of the proposed method, in terms of voiced segments detection, were verified. To reach this aim, the segmentation step was applied on a database consisting of both audio and concurrent electroglottographic (EGG) recordings [27]. Specificity and sensitivity of the proposed method for the detection of voiced segments were estimated, considering the voiced segments as revealed by EGG as the ground truth.

2.2.2 Emotional Speech Database

After the testing phase of the algorithm performances, the described features were firstly estimated on an emotional speech database [28]. Ten different sentences, acted by ten different actors (5 females) playing four different emotions (anger, boredom, happiness and neutral), were selected.

2.2.3. Bipolar patients and healthy subjects

Eleven psychiatric patients (5 females, 40 ± 9) were recruited for this study. All patients had a clinical diagnosis of bipolar disorder. Namely they fulfil the criteria for one of the following mood episode: depressive, mixed, hypomanic. Patients were examined with the structured clinical interview for DSM-IV-TR (SCID) [29]. Clinical rating scales were used to evaluate the presence and the severity of mood symptoms. Particularly the Quick Inventory of Depressive Symptomatology – Clinician Rating (QIDS-C) [30] was used to assess depressive symptoms and the Young Mania Rating Scale (YMRS) [31] for the manic ones. Both scales were administered by a trained physician or psychologist. In this study four different mood states were identified, namely depressed, euthymic, mixed state and hypomanic state. Mood state was labelled according to the scores of the above mentioned clinical rating scales. Particularly QIDS-C higher than eight indicates depressive state, YMRS higher than six indicates (hypo)manic state. Mixed state is labelled if both the scales were over the cut-off while if the patient scores under the cut-off in both of the scales he or she is labelled as euthymic.
We also recruited 18 healthy control subjects (9 females, 30 ± 5). Healthy control subjects did not refer any actual or past psychiatric disorder, and have no history of neurological or major somatic conditions. At the moment of the study they were not taking any medication.

The experimental protocol, approved by the clinical ethical committee, consisted of two different tasks:

- neutral text reading: subjects were asked to read a text that was supposed not to elicit a strong emotional reaction;
- commenting of TAT (Thematic Apperception Test) images: subjects were asked to comment a series of TAT images [32]. The images in this task represent social situations, and they require a personal interpretation of the scene by the subject.

The signals were acquired with a sampling frequency equal to 48 KHz and a resolution of 32 bits by means of a high quality microphone (AKG Perception P220 Condenser Microphone, M-Audio Fast-Track). The mean recording length was 220 s for neutral text reading and 350 s for TAT. The recording sessions for all subjects took place in the afternoon. Each subject was recorded twice in two different days. For seven of the eleven patients one additional recording session took place in the morning of the same day when the afternoon recording was performed. The additional recording consisted in a neutral text reading task.
Control subjects were asked to perform the same tasks of the patients for a total of 18 neutral text double recording sessions and 10 commenting of TAT double recording sessions.

2.3 Statistical Analysis

As regards the emotional speech database, statistical tests were performed to evaluate differences among the features estimated from speeches classified as having a different emotional content. The tests were performed both at single subject level (i.e. intra-subject) and at group level (i.e. inter-subject).

As regards bipolar patients data, intra-subject analyses were performed to test for statistically significant features changes between records related to different mood states. The comparison was only performed between feature sets belonging to the same task. The limited number of subjects with the same mood labels did not allow performing a reliable inter-subject analysis.

To test the specificity of the proposed features with respect to mood changes, features estimated from different recording sessions showing the same labels were compared. This was accomplished by comparing morning and afternoon recording sessions from bipolar patients, and by comparing acquisitions performed on control subjects, which were in the euthymic state in all the recordings.

Parametric and non parametric statistical tests were used according to feature distributions. Gaussianity of feature distribution was tested using a Lilliefors test. The non parametric statistical tests were the Mann-Whitney U-test for intra-subjects analysis and the Kruskal-Wallis test for inter-subject analysis. The parametric test employed was 1-way ANOVA.

3. Results

3.1 Segmentation results

The 94% of audio signal labelled as voiced by the proposed method is found to be voiced according to EGG signal segmentation. On the other side, the 77% of EGG signal labelled as voiced obtains the same classification by using the proposed method on audio records. In conclusion, the proposed approach results in a specificity of 90% and a sensitivity of 81%. Since the percentage of detected vowels by our approach is equal to 95.3, the above-described results can be partially explained by an underestimation of voiced segments length.

3.2 Features statistical distribution

All features estimated at single subject level are not normally distributed after performing a Lilliefors test, so the Mann-Whitney U-test is used for the intra-subject analysis. Regarding the inter-subjects analysis a Kruskal-Wallis test is used to test possible differences among conditions in ampl*, dur* and tilt*. Instead a one-way ANOVA is used with PosSlope, AbsNegSlope, SumDer and GlobalSlope since at group level they are normally distributed.

3.3 Emotional Database results

Both intra subject analysis (data not shown) and inter subject analysis show statistically significant differences among emotions characterized by high arousal with respect to lower arousal states (happiness and anger vs. boredom and neutral). Moreover, regarding intra-subject analysis, in some subjects differences are seen between neutral and boredom, while no differences are displayed between anger and happiness.

The inter-subject analysis reveals that ampl* and tilt* allow to separate anger and happiness from boredom and neutral. Dur* instead allows separating boredom from happiness and anger. In Fig. 3, 4 and 5 the results of the Kruskal-Wallis test are shown. In the graphs each group mean-ranks are represented by a symbol and an interval around the symbol. If two intervals are disjoint, the groups are significantly different. If their intervals overlap, they are not significantly different. Likewise the graphs related to the one-way ANOVA applied to PosSlope, AbsNegSlope, SumDer and GlobalSlope are interpretable. In fact PosSlope (Fig. 6), AbsNegSlope (Fig. 7), SumDer (Fig. 8) and GlobalSlope (Fig. 9) show to be able to separate anger and happiness from boredom and neutral. In table 1 the p-values resulting from the tests and the median or mean of each group are reported. The emotional speeches characterized by a lower arousal resulted in features with a median value lower than that obtained from recordings the emotional speeches whose arousal level is higher.

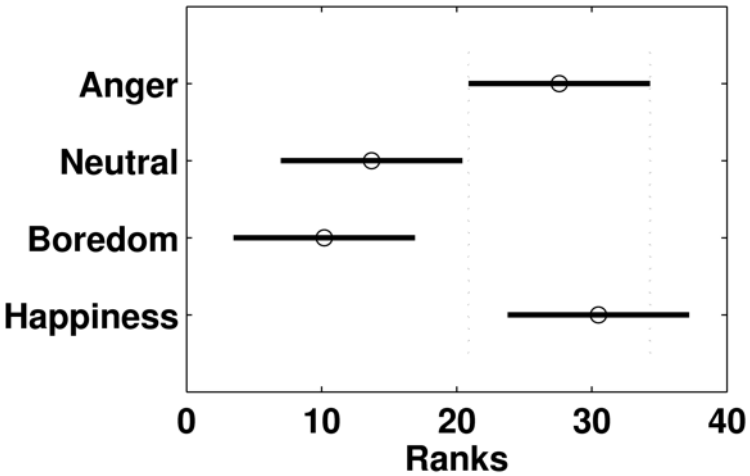

Figure 3: results at group level of emotional speech data. Graphs of Kruskal-Wallis test of Ampl*.

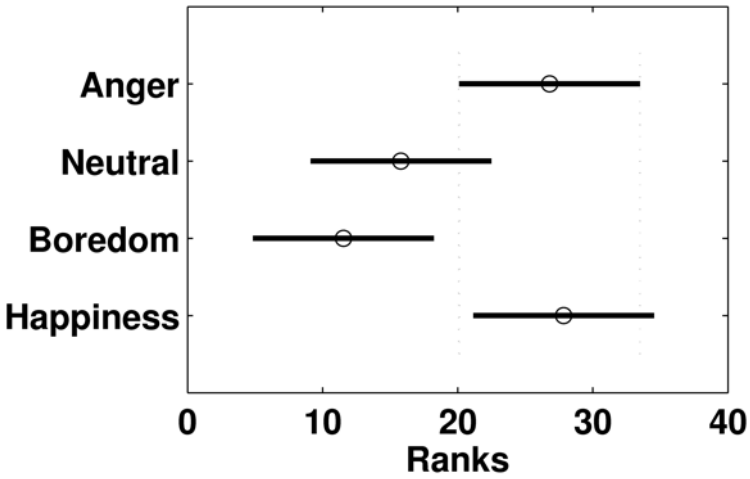

Figure 4: results at group level of emotional speech data. Graphs of Kruskal-Wallis test of Dur*.

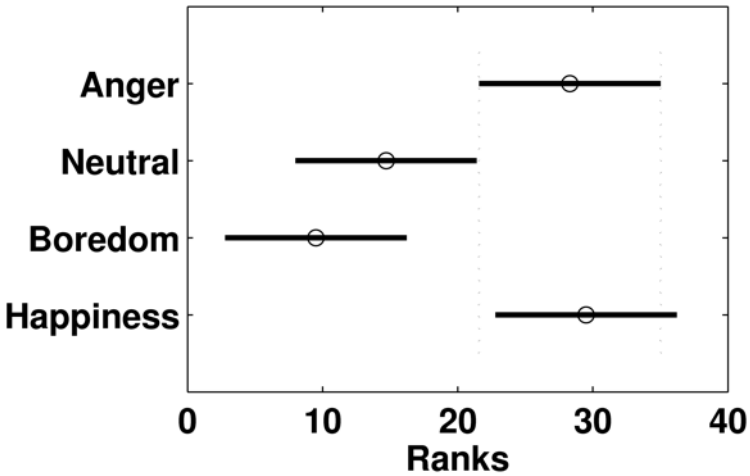

Figure 5: results at group level of emotional speech data. Graphs of Kruskal-Wallis test of Tilt*.

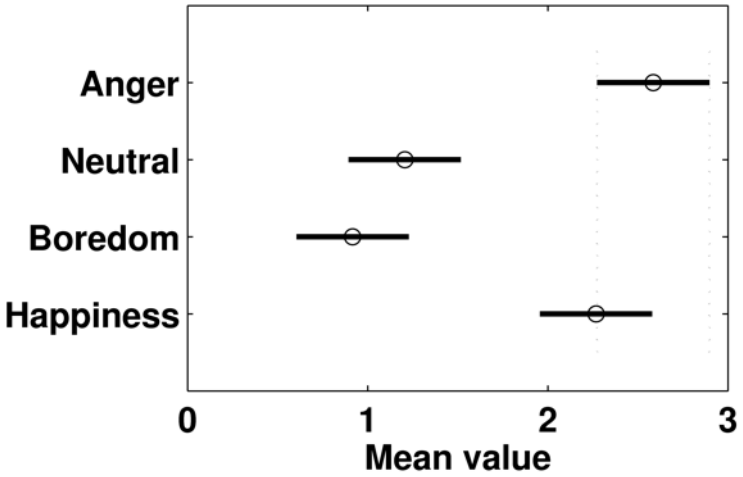

Figure 6: results at group level of emotional speech data. Graphs of one-way ANOVA test of PosSlope.

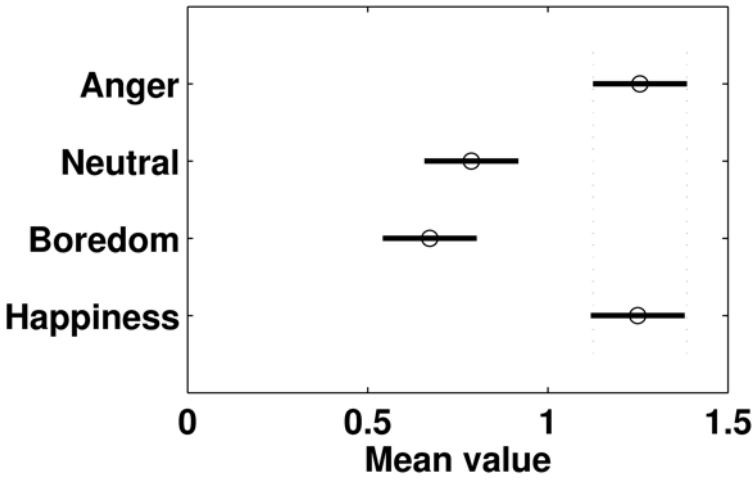

Figure 7: results at group level of emotional speech data. Graphs of one-way ANOVA test of AbsNegSlope.

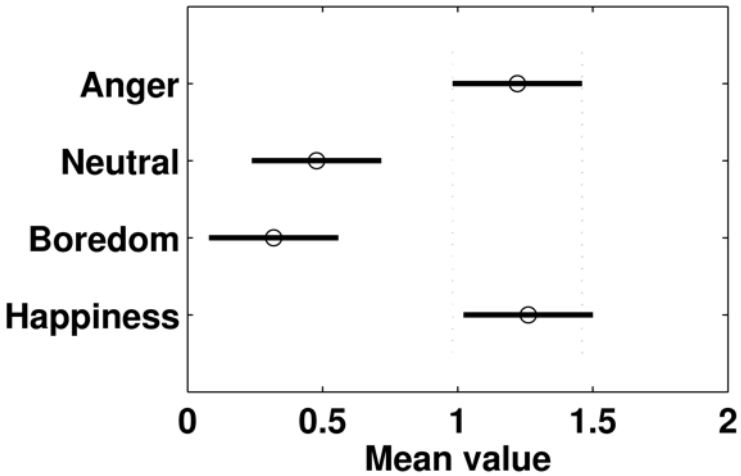

Figure 8: results at group level of emotional speech data. Graphs of one-way ANOVA test of SumDer.

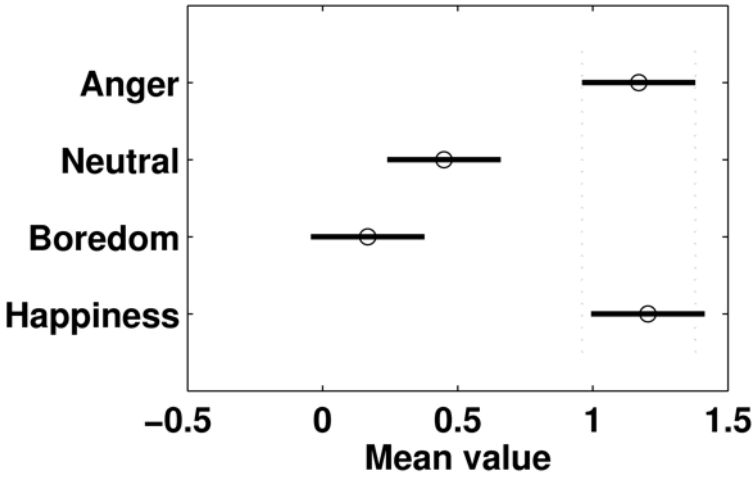

Figure 9: results at group level of emotional speech data. Graphs of one-way ANOVA test of GlobalSlope.

		median			
feature	**p-value**	**anger**	**neutral**	**boredom**	**happiness**
ampl*	6,07E-05	0,47	-0,01	-0,16	0,56
dur*	2,39E-03	-0,01	-0,27	-0,39	0,05
tilt*	7,55E-05	0,18	-0,15	-0,25	0,25
PosSlope	8,96E-09	2,58	1,2	0,91	2,27
AbsNehSlope	8,72E-08	1,25	0,78	0,67	1,24
SumDer	1,46E-06	1,22	0,47	0,31	1,26
GlobalSlope	2,40E-08	1,17	0,44	0,16	1,2

Table 1
Inter-subject analysis results of emotional speeches.
Median values and p-values of the Kruskal-Wallis tests are shown

3.4 Bipolar patients results

In table 2 the classification of bipolar patients performed by clinicians is shown. All the subjects show a different mood state in the second acquisition day with respect to the first one. Six patients out of eleven display as first state the depressed state. Nine patients, instead, are labelled in the euthymic state in the second recording day.

	day 1	**day 2**
A	Hypomania	Euthymia
B	Hypomania	Euthymia
C	Hypomania	Euthymia
D	Depression	Euthymia
E	Depression	Euthymia
F	Depression	Hypomania
G	Hypomania	Euthymia
H	Depression	Euthymia
I	Depression	Euthymia
L	Depression	Euthymia
M	Mixed	Depression

Table 2
Patients' mood states in each day

Concerning the analysis of audio signals acquired in subjects experiencing a different mood state, an intra-subject analysis was performed. Only the features related to the same task category were compared. With regard to the neutral text reading, seven patients out of eleven show statistically significant differences between ampl* features (table 3). In particular, in three patients out of four the median value was found to decrease passing from hypomania to euthymia. Ampl* value in hypomanic state was also found to be higher than the value estimated from depressive state. In two out of five patients passing from depression to euthymia, an increase in ampl* median value is observed; while in the other three subjects no statistically significant differences are observed. Analysis on the commenting of TAT images showed that ampl* features highlights statistically significant differences in five patients out of eleven. In these case ampl* feature median values decreases in two patients passing from depression to euthymic state, thus showing for these patients an opposite trend with respect to neutral text reading. Comparing the results obtained in this task with respect to neutral text reading, fewer differences were highlighted when hypomanic state was observed. However, when differences were found, the ampl* values in hypomanic state were higher.

			READING			TAT		
	Mood state		Amplitude*			Amplitude*		
	day 1	**day 2**	**day 1**	**day 2**	**p-value**	**day 1**	**day 2**	**p-value**
A	Hyp.	Eut.	**0.12 [0.84]***	**-0.36 [0.63]***	**3.70E-03***	0.06 [0.72]	-0.05 [0.89]	3.04E-01
B	Hyp.	Eut.	**-0.34 [0.66]***	**-0.57 [0.42]***	**3.92E-02***	**-0.03 [0.88]+**	**-0.22 [0.77]+**	**4.26E-02+**
C	Hyp.	Eut.	**-0.45 [0.54]***	**-0.20 [0.80]***	**3.67E-02***	-0.33 [0.66]	-0.38 [0.61]	3.51E-01
D	Dep.	Eut.	-0.50 [0.49]	-0.59 [0.40]	9.38E-01	**-0.05 [0.94]+**	**-0.37 [0.62]+**	**1.41E-03+**
E	Dep.	Eut.	**-0.43 [0.57]***	**-0.04 [0.95]***	**2.43E-02***	**0.36 [0.55]+**	**0.03 [0.96]+**	**5.62E-03+**
F	Dep.	Hyp.	**0.32 [0.62]***	**0.47 [0.50]***	**3.99E-02***	**0.08 [0.72]+**	**0.28 [0.57]+**	**4.23E-03+**
G	Hyp.	Eut.	**0.44 [0.52]***	**0.27 [0.68]***	**2.71E-02***	-0.10 [0.89]	0.00 [0.87]	1.65E-01
H	Dep.	Eut.	0.15 [0.76]	0.06 [0.89]	6.69E-01	0.15 [0.68]	-0.21 [0.78]	4.58E-01
I	Dep.	Eut.	**-0.81 [0.18]***	**-0.25 [0.74]***	**1.55E-04***	-0.07 [0.72]	-0.04 [0.82]	8.01E-01
L	Dep.	Eut.	-0.07 [0.92]	0.12 [0.84]	4.17E-01	0.16 [0.83]	-0.09 [0.90]	4.78E-01
M	Mix.	Dep.	0.12 [0.71]	0.02 [0.62]	2.00E-01	**0.11 [0.64]+**	**-0.05 [0.76]+**	**2.61E-02+**

Table 3
Median and median absolute deviation [mad] of Amplitude* as estimated from bipolar patients.
The symbols (* or +) indicate p-values < 0.05 in Mann-Whitney U-test related to ampl* features.

The results pertaining the other features show statistically significant differences between two days recordings performed with subjects in a different mood state. However, no coherent direction change among subjects experiencing the same mood swing, could be highlighted. Moreover, the changes do not share the same sign across the two tasks. In particular dur* features highlights statistically significant differences in six patients out of eleven in neutral text reading audio signals, while, in signals related to the commenting of TAT images differences were observed in two subjects (data not shown). The analysis on tilt* feature returns five significant p-values in the audio recordings of the neutral text reading, and three p-values in those deriving from the other task (data not shown). In six patients out of eleven PosSlope shows statistically significant differences concerning neutral text reading task and four differences in the TAT task (data not shown). The results pertaining AbsNegSlope (table 4) reports eight significant differences between the two recording days with regard to the neutral text reading task, and six statistically significant differences concerning the commenting of TAT task. As regards neutral text reading, in three patients out of four the euthymic state shows a lower value of AbsNegSlope than the value found in hypomanic state. The same behaviour was found in TAT task. However, only in patients B and C the same change sign was found in both tasks. SumDer features shows statistically significant differences in six patients out of eleven, regarding the neutral text reading, while no differences were observed from recordings of the commenting of TAT images (data not shown). Finally GlobalSlope in neutral text reading reports differences in seven out of eleven subjects (data not shown). In all cases, except two, GlobalSlope value was closer to zero for subjects in euthymic state. In two subjects statistically significant differences were found only in one feature extracted from signals acquired during TAT task. In particular, subject L showed differences only in AbsNegSlope, while as regards subject M significant differences were found in ampl*. Inter-subjects analysis is not performed here since the number of enrolled subjects is too small.

			READING			TAT		
	Mood state		AbsNegSlope			AbsNegSlope		
	day 1	**day 2**	**day 1**	**day 2**	**p-value**	**day 1**	**day 2**	**p-value**
A	Hyp.	Eut.	**0,45 [0,31]***	**0,57 [0,36]***	**1,70E-03***	**0,61 [0,36]+**	**0,49 [0,32]+**	**9,60E-03+**
B	Hyp.	Eut.	**1,28 [0,97]***	**0,95 [0,72]***	**1,07E-02***	**0,57 [0,39]+**	**0,46 [0,31]+**	**3,29E-03+**
C	Hyp.	Eut.	**0,64 [0,34]***	**0,53 [0,31]***	**2,51E-02***	**0,44 [0,29]+**	**0,40 [0,23]+**	**2,22E-02+**
D	Dep.	Eut.	**0,32 [0,20]***	**0,39 [0,26]***	**1,56E-02***	0,38 [0,26]	0,42 [0,29]	5,68E-02
E	Dep.	Eut.	0,36 [0,18]	0,45 [0,32]	5,06E-02	0,43 [0,24]	0,47 [0,27]	2,00E-01
F	Dep.	Hyp.	**0,45 [0,26]***	**0,53 [0,30]***	**1,65E-02***	0,69 [0,60]	1,01 [0,70]	2,16E-01
G	Hyp.	Eut.	**0,69 [0,40]***	**0,61 [0,36]***	**2,10E-02***	**0,55 [0,34]+**	**0,63 [0,41]+**	**1,53E-02+**
H	Dep.	Eut.	**0,59 [0,36]***	**0,46 [0,31]***	**1,80E-02***	**0,46 [0,30]+**	**1,18 [0,67]+**	**6,90E-03+**
I	Dep.	Eut.	**0,65 [0,38]***	**0,49 [0,32]***	**5,39E-04***	0,40 [0,31]	0,48 [0,36]	6,46E-01
L	Dep.	Eut.	0,54 [0,38]	0,49 [0,36]	2,24E-01	**0,33 [0,22]+**	**0,48 [0,35]+**	**2,52E-02+**

M	Mix.	Dep.	0,64 [0,38]	0,58 [0,35]	9,45E-01	0,46 [0,30]	0,39 [0,27]	1,68E-01

Table 4
Median [mad] of AbsNegSlope as estimated from bipolar patients.
The symbols (* or +) indicate p-values < 0.05 in Mann-Whitney U-test related to AbsNegSlope features.

3.4.1 Features Specificity

Intra-subject analyses were applied to data with the same labels to check for the specificity of the proposed features. In order to have a good specificity, the tests in this case should not show any statistically significant difference. Specificity was investigated analysing the data acquired in the same day from bipolar patients and the data acquired from healthy control subjects in different days.
In tables 5 and 6 the tests about possible differences between features acquired from bipolar patients in the morning and in the afternoon sessions of first recording day (day 1) are shown. At day 1, no statistically significant differences were found between Taylor-inspired features across all subjects (table 5). In day 2 a difference was found for subject B in ampl* and for subject E both in dur* and tilt* (data not shown).
In table 6 the results related to the second category of features estimated at day 1 from bipolar subjects are summarized. SumDer and GlobalSlope did not show any statistically significant difference either at day 1 or at day 2 (data not shown). PosSlope revealed two statistically significant differences at day 1 and no differences at day 2. AbsNegSlope did not show any difference at day 1 and only one difference at day 2 for subject G (data not shown).

	ampl*		**dur***		**tilt***	
A	0,05[0,86]	0,12[0,84]	-0,50[0,45]	-0,40[0,59]	-0,42[0,57]	-0,34[0,65]
B	-0,57[0,43]	-0,34[0,66]	-0,69[0,30]	-0,66[0,33]	-0,71[0,28]	-0,62[0,37]
C	-0,45[0,54]	-0,40[0,25]	-0,66[0,33]	-0,63[0,36]	-0,67[0,32]	-0,67[0,32]
D	-0,50[0,49]	-0,59[0,40]	-0,50[0,46]	-0,50[0,45]	-0,31[0,68]	-0,47[0,52]
E	-0,43[0,57]	-0,40[0,95]	-0,75[0,25]	-0,77[0,22]	-0,75[0,24]	-0,78[0,21]
F	0,40[0,54]	0,32[0,62]	-0,33[0,66]	-0,30[0,69]	-0,05[0,94]	-0,09[0,90]
G	0,38[0,57]	0,44[0,52]	-0,33[0,66]	-0,36[0,63]	-0,16[0,83]	-0,16[0,83]

Table 5
Results of day 1 sessions concerning ampl*, dur* and tilt* features as estimated from bipolar patients.
Median and Mad (in square brackets) values are shown. No statistically significant differences were found.

	SumDer		**GlobalSlope**		**PosSlope**		**AbsNegSlope**	
A	0,16[0,36]	0,17[0,34]	-0,05[0,66]	-0,01[0,69]	0,54[0,32]	0,58[0,35]	0,44[0,29]	0,45[0,31]
B	0,10[0,61]	0,17[0,71]	-0,40[1,42]	-0,44[1,50]	0,85[0,57]	0,83[0,52]	1,00[0,75]	1,28[0,97]
C	0,12[0,44]	0,13[0,40]	-0,27[0,95]	-0,25[0,97]	0,70[0,43]	0,75[0,39]	0,64[0,34]	0,60[0,36]
D	0,18[0,30]	0,16[0,22]	-0,02[0,64]	-0,03[0,57]	**0,52[0,23]***	**0,38[0,19]***	0,32[0,20]	0,39[0,24]
E	0,07[0,27]	0,15[0,33]	-0,22[0,62]	-0,24[0,62]	**0,45[0,26]***	**0,59[0,31]***	0,36[0,18]	0,36[0,20]
F	0,33[0,42]	0,40[0,48]	0,15[0,71]	0,17[0,66]	0,78[0,43]	0,79[0,40]	0,47[0,30]	0,45[0,26]
G	0,47[0,70]	0,30[0,63]	0,13[0,92]	0,20[0,88]	1,11[0,63]	0,99[0,60]	0,73[0,45]	0,69[0,40]

Table 6
Results coming from the analysis on day 1 data concerning SumDer, GlobalSlope, PosSlope and AbsNegSlope features as estimated from bipolar patients. The symbol * indicates p-values < 0.05 in Mann-Whitney U-test related to patients' features.

Regarding the tests on control subjects (data not shown) performing the neutral text reading task, ampl*, dur* tilt* and GlobalSlope did not show any statistically significant differences between different days. PosSlope, AbsNegSlope and SumDer instead showed statistically significant differences respectively in 3, 4 and 2 subjects out of 18. Analyzing the data about TAT images commenting, ampl*, tilt*, AbsNegSlope and GlobalSlope did not show any statistically significant difference, while on the contrary dur* and AbsNegSlope reported a significant difference in 1 out of 10 subjects, PosSlope in 3 out of 10, and SumDer in 2 subjects out of 10.

4. Discussion

In this work, possible changes in speech related features were investigated, to discriminate between different emotions in speech and between different mood states in bipolar subjects. In particular we investigated the use of parameters describing fundamental frequency changes in voiced part of syllables.

One important step of the proposed approach is related to a correct segmentation of the sentences. The specificity of this step, as assessed using EGG data, is found to be good while a lower sensitivity was found. We believe that the segmentation results obtained are good since we are more interested in a low probability of labelling unvoiced segments as voiced.
The good performances of the swipe' for F0 estimation, were already described in literature by comparing its results with other state of the art algorithms by Camacho [25] and Evanini [33]. In [34] swipe' algorithm was used to estimate F0 and a jitter-related measure on voiced segments, and its performances were compared with those achievable with the SIFT algorithm. The performances of the two approaches were similar as concerns average F0 on each voiced segment. On the other hand, swipe' was found to outperform SIFT as regards jitter estimation.
In this work we propose two categories of features. The first one is inspired by those introduced by the Taylor's Tilt Intonational model. We have to point out that the features here proposed are only functionally equivalent to Taylor's ones. In fact, the former are estimated from all voiced segments of syllables while the tilt model takes into account intonational events. The detection of these events is usually performed with a classifier that requires hand labelled sentences by a human. The approach we propose is simpler and completely automatic.
The second kind of features instead is correlated to the speed of variation of F0 as estimated from the F0 contour.
Analysis on the emotional speech database demonstrated that the proposed parameters allow highlighting significant differences among different emotional speech recordings. Such differences were observed both in intra- and in inter- subject analysis. In particular, concerning inter-subject analysis, some features have been shown to be capable of grouping emotions by excitation level. The more the subjects are aroused, the more their speech features exhibit differences. These results seem to be in agreement with Pakosz, who sustained that intonation can only carry information about the level of emotional arousal [35]. Banse and Scherer [36] showed how arousal has a powerful effect on vocal expressions often confounding the effects of valence or potency/control [37]. However, it is important to stress that the emotional database we took into account is a collection of sentences spoken by actors who were "playing" different emotions, while the actors' actual mood is unknown. Vogt and André [16] showed the need of partially overlapped features sets to recognize different emotions from acted and spontaneous speeches. Anyway, Bänziger and Scherer [19] defended in a detailed way the prudent use of acted sentences in the study of emotion. In fact, the difficulties to record different and often rare emotional states from the same subjects, and to assess each emotional state might allow acted speeches datasets to give an important contribution in this field. Schuller et al. [14] sustained that acted corpora have two disadvantages: the first one is that acting emotions is different from producing "spontaneous" emotions [38] and secondly, the prompted types of emotions are not the same as those in realistic scenarios. So while the acquisition of realistic corpora is envisaged, using acted corpora could be convenient for benchmarking, even if the relationship between the results coming from the two kind of dataset is unclear [39]. In our view, the results on the emotional dataset we obtained could be important to evaluate the capability of the implemented algorithm to extract prosodic features that can be used in a tool to estimate subjects' mood state. The analysis of real and not acted emotional speeches could then give the final validation to the proposed features.
For these reasons, in the present work, we do not infer the goodness of the results on bipolar patients from those on the emotional database. In fact, the pathophysiological factors influencing speech in bipolar disorders could lead to completely different phenomena linking subjects' mood states to voice production. For this purpose, repeated acquisitions on bipolar patients, both in controlled and real life scenarios, will be important for clarifying the relevance of the features we investigated.
Intra-subject analyses on bipolar patients have shown that the proposed features have a good specificity. In almost every comparison between features extracted from acquisitions labelled with the same mood state, no statistically significant differences are found out. To test for specificity in bipolar patients, double recording sessions were performed in the same day. To improve the statistical significance of the specificity value, additional recordings with the same label but acquired in different days could be used. Very good results were found by analysing data acquired from healthy subjects in different days. In particular Taylor-inspired features and GlobalSlope demonstrated very high specificity. The remaining features showed a good specificity as applied to neutral text reading, while worst results were found from TAT recordings.
As a result, most of statistically significant differences are found between features estimated when patients are scored in a different mood state.

Overall, this study shows that the direction of the change is not coherent across subjects. Only ampl* seems to have a more coherent behaviour across subjects. In particular, when statistically significant differences were found out, the feature values extracted from recordings related to the hypomanic state are demonstrated to be higher than the other ones in every subject, but one, irrespective of the task. We have to point out when one of the two states was the hypomanic state, a difference was always found out as regards the neutral text reading task. The same behaviour was no observed by analysing TAT recordings.

A study involving a larger number of subjects could reveal potential predominant behaviour of the proposed features. However, other factors could influence this subject-dependent behaviour, e.g. subject anxiety level. In [7] a similar phenomenon was observed and the differences were hypothesized to be caused by anxiety. The same subject-specific behaviour was observed in [26] were F0 average level, jitter and F0 standard deviation were considered. As a possible confirmation of this hypothesis, in [40] the authors suggest that some vocal parameters, for example F0, can be used as objective markers of Social Anxiety.
Anxiety level could also be a factor affecting specificity results, even though we cannot exclude the relevance of other unobserved factors. In particular, as regards the features extracted from bipolar patients during double recording sessions, these factors could include boredom or fatigue for a full day visit.

No statistical analyses of intra tasks results were performed. However, the directions of feature changes are not the same across the two tasks. A task dependent behaviour of F0 has also been observed by Horwitz et al. [10] in depressed patients where the correlations between F0 and a score of depression were investigated. The tasks were the reading of a text and sample from conversational speech. The phenomenon we reported here could be partially explained, taking into account the differences between the two tasks. In fact the description of the images of the TAT, involves complex brain processes requiring interpretation of the images.

Possible improvements of the proposed work would be the inclusion of other features, as those related to glottal, spectral and energy parameters, as well as a more detailed description of the temporal dynamics of F0 at sentence and word level [18].

5. Conclusions

A method to estimate prosodic features of voiced segments is proposed in this work. Two categories of features are investigated. The first category is comprised of features that are functionally equivalent to those introduced by the Taylor's model. Otherwise than what indicated by the latter model, here those features are estimated on all the voiced events. This choice allows an easier approach but it results in features with a different meaning. The other category of features is chosen to describe the speed of F0 change. Statistically analysis on features estimated from an emotional speech database reveals how the proposed features could be able to highlight significant differences among the recordings. In particular emotional speeches can be classified according to the excitation level of the acted emotion.

The analyses on bipolar patients and controls have highlighted that Taylor-inspired features and GlobalSlope have a very good specificity. In fact, statistically significant differences have not been detected in almost all the couples of records characterized by the same label. Statistically significant differences have been found out comparing speech records corresponding to different mood states.

The main contributions of this work are that the directions of change estimated for different patients experiencing the same mood swing, are not coherent and that the changes seem to be task dependent. Only ampl* seems to have a coherent behaviour among patients experiencing a mood swing from or to hypomania. An increase in the number of enrolled patients, as well as the analysis of more time points per subject, would be useful for highlighting possible predominant trends of the proposed features with respect to mood changes.

However, we believe that the results here described furnish helpful indications for planning further investigations. Moreover, they indicate that a model of speech changes might be subject-specific and that a richer characterization of subject status, for instance adding the anxiety dimension in the descriptors, could explain part of the observed variability. In conclusion, we believe that the proposed features might give a relevant contribution to the demanding research field of speech-based mood classifiers.

Acknowledgements

This is a study-part of the European project PSYCHE (Personalised monitoring SYstems for Care in mental HEalth) funded by the Seventh Framework Programme, ICT-2007.5.1.

References

[1] G. Valenza, C. Gentili, A. Lanata, E. P. Scilingo, Mood recognition in bipolar patients through the psyche platform: preliminary evaluations and perspectives, Artif. Intell. Med. 57 (1) (2013) 49–58.
[2] A. Greco, A. Lanata, G. Valenza, G. Rota, N. Vanello, E. Scilingo, On the deconvolution analysis of electrodermal activity in bipolar patients, In EMBS (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, 6691–6694.
[3] G. Valenza, M. Nardelli, A. Lanata, C. Gentili, G. Bertschy, R. Paradiso, E. P. Scilingo, Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis, IEEE J. Biomed. Health Inform. 18 (5) (2014) 1625–1635.
[4] C. Sobin, H. A. Sackeim, Psychomotor symptoms of depression, Am. J. Psychiat. 154 (1) (1997) 4–17.
[5] Å. Nilsonne, J. Sundberg, S. Ternstrom, A. Askenfelt, Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression, J. Acoust. Soc. Am. 83 (1988) 716-728.
[6] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, P. J. Snyder, Voice acoustical measurement of the severity of major depression, Brain Cognition, 56 (1) (2004) 30–35.
[7] E. Moore, M. A. Clements, J. W. Peifer, & L. Weisser, Critical analysis of the impact of glottal features in the classification of clinical depression in speech, IEEE Trans. Bio. Med. Eng., 55(1), (2008) 96-107.
[8] L.S. Low, M.C. Maddage, M. Lech, L.B. Sheeber, N.B. Allen. Detection of clinical depression in adolescents' speech during family interactions. IEEE T. Bio-Med. Eng., 58(3) (2011) 574-586.
[9] K.E.B. Ooi, M. Lech, N.B. Allen. Multichannel weighted speech classification system for prediction of major depression in adolescents. IEEE T. Bio-Med. Eng., 60(2) (2013) 497-506
[10] R. Horwitz, T. F. Quartieri, B. S. Helfer, B. Yu, J. R. Williamson, J. Mundt. On the Relative Importance Vocal Source, System, and Prosody in Human Depression, Body Sensor Networks (BSN), 2013 IEEE International Conference on, 6-9 May 2013, 1-6
[11] S. G. Koolagudi, K. S. Rao, Emotion recognition from speech: a review, International Journal of Speech Technology 15 (2) (2012) 99–117.
[12] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recogn. 44 (3) (2011) 572–587.
[13] K. R. Scherer, Vocal affect expression: a review and a model for future research, PSYCHOL BULL 99 (2) (1986) 143.
[14] B. Schuller, A. Batliner, S. Steidl, D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech. Commun., 53 (9) (2011) 1062-1087.
[15] T. Vogt, E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. (2005) 474-477.

[16] M. Bulut, S. Narayanan, On the robustness of overall f0-only modifications to the perception of emotions in speech, J. Acoust. Soc. Am. 123 (2008) 4547-4558.
[17] J.P. Arias, C. Busso, N.B. Yoma Shape-based modeling of the fundamental frequency contour for emotion detection in speech. Comp. Speech Lang. 28 (2014) 278-294
[18] K. Rao, S. Koolagudi, R. Vempada, Emotion recognition from speech using global and local prosodic features, International Journal of Speech Technology 16 (2) (2013) 143–160.
[19] T. Banziger, K. Scherer, The role of intonation in emotional expressions, Speech Commun. 46 (3) (2005) 252–267.
[20] H. Wang, A. Li, Q. Fang, F0 contour of prosodic word in happy speech of mandarin, in: Active Computing and Intelligent Interaction, Springer (2005) 433–440.
[21] M. Lin, On production and perception of boundary tone in chinese intonation, in: International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, (2004) 125-130
[22] P. Taylor, Analysis and synthesis of intonation using the tilt model, J. Acoust. Soc. Am. 107 (2000) 1697-1714.
[23] B. Atal, L. Rabiner, A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, IEEE T. Acoust. Speecj, 24 (3) (1976) 201–212.
[24] N. H. de Jong, T. Wempe, Praat script to detect syllable nuclei and measure speech rate automatically, Behav. Res. Methods. 41 (2) (2009) 385–390.
[25] A. Camacho, J. G. Harris, A sawtooth waveform inspired pitch estimator for speech and music, J. Acoust. Soc. Am. 124 (2008) 1638-1652.
[26] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanata, E. P. Scilingo, Speech analysis for mood state characterization in bipolar patients, In EMBS (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, 2104–2107.
[27] J. Kominek, A. Black, CMU ARCTIC databases for speech synthesis CMU Language Technologies Institute, Language Technologies Institute, CMU, Pittsburgh PA, Tech Report CMU-LTI-03-177.
[28] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, A database of german emotional speech., in: Interspeech, (2005) 1517–1520.
[29] M. B. First, R. L. Spitzer, M. Gibbon, J. B. W. Williams. Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition. (SCID-I/P) New York: Biometrics Research, New York State Psychiatric Institute (2002).
[30] A. J. Rush, M.H. Trivedi, H.M. Ibrahim, T.J. Carmody, B. Arnow, D. N. Klein, J.C. Markowitz, P.T. Ninan, S. Kornstein, R. Manber, M.E. Thase, J.H. Kocsis, M.B. Keller. The 16-item Quick Inventory of Depressive Symptomatology (QIDS) Clinician Rating (QIDS-C) and Self-Report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. Biol. Psychiat. 54 (2003) 573-583.
[31] R.C. Young, J.T. Biggs, V.E. Zeigler, D.A. Meyer. A rating scale for mania: reliability, validity and sensitivity. Brit. J. Psychiat., 133 (1978) 429–435.
[32] H. A. Murray, Uses of the thematic apperception test, Am. J. Psychiat., 107 (8) (1951): 577-581.
[33] K. Evanini, C. Lai, The importance of optimal parameter setting for pitch extraction, J. Acoust. Soc. Am. 128 (2010) 2291-2291
[34] N. Vanello, N. Martini, M. Milanesi, H. Keiser, M. Calisti, L. Bocchi, C. Manfredi, L. Landini, Evaluation of a pitch estimation algorithm for speech emotion recognition, in: Proc. 6th MAVEBA, 2009, 29–32.
[35] M. Pakosz, Attitudinal judgments in intonation: Some evidence for a theory, J. Psycholinguist. Res. 12 (3) (1983) 311–326.
[36] R. Banse, K. R. Scherer, Acoustic profiles in vocal emotion expression., J. Pers. Soc. Psychol. 70 (3) (1996) 614.
[37] M. Goudbeek, K. Scherer, Beyond arousal: Valence and potency/control cues in the vocal expression of emotion, J. Acoust. Soc. 128 (3) (2010) 1322–1336.
[38] D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, Y. Shibuya, Exploratory study of some acoustic and articulatory characteristics of sad speech. Phonetica, 63(1) (2006) 1-25
[39] B. Schuller, A. Batliner. Computational paralinguistics: emotion, affect and personality in speech and language processing, Taxonomies, John Wiley & Sons, (2013) 21-53.
[40] E. Gilboa-Schechtman, L. Galili, Y. Sahar, O. Amir. Being “in” or “out” of the game: subjective and acoustic reactions to exclusion and popularity in social anxiety, Frontiers in Human Neuroscience (2014) 8