34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Speech analysis for mood state characterization in bipolar patients

Nicola Vanello[1], Andrea Guidi[2], Claudio Gentili[3], Sandra Werner[4],
Gilles Bertschy[5], Gaetano Valenza[2], Antonio Lanatá[2] and Enzo Pasquale Scilingo[2]

***Abstract*—Bipolar disorders are characterized by an unpredictable behavior, resulting in depressive, hypomanic or manic episodes alternating with euthymic states. A multi-parametric approach can be followed to estimate mood states by integrating information coming from different physiological signals and from the analysis of voice. In this work we propose an algorithm to estimate speech features from running speech with the aim of characterizing the mood state in bipolar patients. This algorithm is based on an automatic segmentation of speech signals to detect voiced segments, and on a spectral matching approach to estimate pitch and pitch changes. In particular average pitch, jitter and pitch standard deviation within each voiced segment, are estimated. The performances of the algorithm are evaluated on a speech database, which includes an electroglottographic signal. A preliminary analysis on subjects affected by bipolar disorders is performed and results are discussed.**

I. Introduction

Bipolar or manic-depressive psychiatric disorder is characterized by cyclic variations of mood status. Subjects may experience hypomania or depression, passing through euthymic state [1]. Currently, clinicians estimate mood state using interviews and questionaries. There is a need for the clinicians to rely upon a decision support system that may help them to improve diagnosis and modulate therapy. Such a system could be used to monitor subjects' mood during every day activity, even outside the clinical setting and could exploit the information coming from multi-parametric acquisitions. Speech related features are good candidates to be included in such a system. In fact, several studies have investigated the characteristics of speech in people with psychiatric disorders, as depression. Prosodic and spectral features were reported to vary in patients with respect to healthy subjects [2]–[6]. In particular glottal source parameters, as those related to pitch, has been found to correlate with mood changes, as supported by studies on depressed and normal subjects [6]. Although with some discrepancies, pitch changes have been observed in depressed patients [3] with respect to controls. More consistent behavior was shown by pitch variability-related features. In fact a decrease of pitch standard deviation has been related to an increase in depression severity [5] [7]. Another measure of pitch variability is jitter that is defined as period to period changes of pitch. While pitch standard deviation is related to long-term changes of pitch, jitter reflects short term changes. Jitter has been proposed as an important feature for the characterization of mood states, since it may reflect a dysregulation of autonomous nervous system that influence muscular tone and articulatory control [8]. In this work we aim at studying glottal features as extracted from running speech to highlight possible changes in mood state in bipolar patients. Average pitch, jitter and pitch standard deviation (pitch SD) will be estimated from voiced segments. An algorithm based on a spectral matching approach [9] for pitch estimation will be described. The performances of the proposed approach will be assessed by applying the algorithm to a speech database containing concurrent acquisitions of electroglottographic (EGG) signals. Preliminary results on speech signals, acquired on bipolar patients in a clinical setting, will be described. Patients were enrolled for a study-part of the European project PSYCHE (Personalised monitoring SYstems for Care in mental HEalth), which is funded by the Seventh Framework Programme.

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

[1]Nicola Vanello is with the Department of Information Engineering, University of Pisa, Pisa, Italy (email: nicola.vanello@iet.unipi.it)

[2]Andrea Guidi, Gaetano Valenza, Antonio Lanatá and Enzo Pasquale Scilingo are with the Interdepartmental Research Center "E. Piaggio", University of Pisa, Pisa, Italy (email: andrea.guidi@for.unipi.it; g.valenza@ieee.org; a.lanata@centropiaggio.unipi.it; e.scilingo@centropiaggio.unipi.it)

[3]Claudio Gentili is with the Deptartment of Psychiatry, Neurobiology, Pharmacology and Biotechnologies, University of Pisa, Pisa, Italy (email: c.gentili@med.unipi.it)

[4]Sandra Werner is with FORENAP, Rouffach, France (email: Sandra.WERNER@forenap.com)

[5]Gilles Bertschy is with the Department of Psychiatry, University Hospital and University of Strasbourg, INSERM u666, France (email: gilles.bertschy@chru-strasbourg.fr)

II. Materials and method

The work here presented can be divided in three phases: the first one is focused on testing the implemented features extraction method on a speech database, the second aims at applying the proposed algorithm on patients' audio recordings and the third phase is concerned on evaluating possible differences in the obtained features across different mood states.

A. Algorithm Description

The proposed algorithm, in a first step, performs a segmentation of speech signal to identify voiced segments. In a second step, the pitch contour in each voiced segment is estimated thus allowing to extract average pitch, jitter and pitch SD. Voiced sounds are characterized by high signal intensity values and by lower frequency components with respect to unvoiced sounds. Signal intensity across time is estimated using the autocorrelation method as applied to a sliding window of length of 80 ms with a time step of 8 ms. The intensity value is obtained by retaining only the frequencies between 5 Hz and 5 kHz. The median of the speech intensity is used as threshold to discard unvoiced segments. To detect syllables nuclei, local maxima and local dips of the obtained intensity contour are analyzed. Using an approach similar to [10], we consider a syllable nucleus to be centered around a local maxima whose intensity was 1 dB higher than the intensity of a preceding

local dip. To discriminate between voiced and possible high intensity unvoiced sound, we calculate signal zero crossing rate (ZCR) [11]. Only the segments that present a low ZCR value are considered as voiced. Pitch, jitter and pitch SD are estimated for each segment by using a sliding window approach and a two-step procedure, the first being an estimate (f_0) of the pitch. The time length of the window and the time shift are calculated in a first step as $T = 4/f_0$ and $dt = T/4$ respectively. In a second step, final pitch estimate is obtained as the mean over the speech segment, while jitter is estimated according to the following formula:

$$Jitter = \frac{1}{N-1}\sum_{i=1}^{N-1}|F_{i+1} - F_i| / \frac{1}{N}\sum_{i=1}^{N} F_i \qquad (1)$$

where F_i is the estimated pitch at the $i-th$ window. This measure results in a under-estimate of the actual jitter, since each pitch value, F_i, is obtained by analyzing a window containing approximately 4 pitch periods. Results of the proposed approach on simulated vowels are reported elsewhere [12]. All the pitch values found within each voiced segment, are used to estimate pitch SD for that segment. Pitch in each time window is estimated using the SWIPE' algorithm [9], as the fundamental frequency of the sawtooth waveform whose spectrum best matches the spectrum of the input signal.

B. Tests on a Speech Database

The performances or the proposed method are evaluated on the CMU Arctic Database [13] containing audio and EGG acquisitions. EGG signal is related to impedance changes due to vocal folds closure thus allowing to estimate pitch and pitch changes reliably. The proposed approach will be evaluated by comparing the features extracted from EGG with those obtained from the audio signal. To estimate pitch and pitch changes from the EGG signal a 5-coefficient-Daubechies wavelet-based filtering operation was performed to deprive the signal of low frequency drifts. On the quasi-periodical detrended signal, a waveform matching algorithm is used to detect glottal cycle timings. This procedure is performed on each voiced segment, using a segment-specific average waveform. The speech corpus we considered consists of approximately 1100 short phrases. High quality audio recordings were sampled using 32 kHz sampling rate. A linear regression model between actual, i.e. obtained from EGG, and estimated features was applied. Since our approach estimates pitch within a moving window of length equal to $T = 4/f_0$ with a time step equal to $dt = T/4$, while pitch values estimated from EGG are estimated on a period-by-period basis, a smoothing operation in time domain was performed on the latter values to allow a proper comparison. In particular final EGG-derived pitch estimates at the *i*-th time step, were obtained as the average of four pitch values.

C. Experimental Protocol

In this preliminary work six psychiatric patients (1 female) were recruited. All subjects had a clinical diagnosis of bipolar disorder, had the competence to lead independent and active lives and had no substance use disorders. They did not show suicidal tendencies and were in absence of delusions or hallucinations. Subjects were recorded in two different days. Before each session, a physician labeled patient's mood status by clinician administered rating scales. In this work three different states were identified: depressed, euthymic and hypomanic states. In each day the experimental protocol, which received hospital ethics committee approval, consisted of two sub-sessions, organized as follows:

- TAT (Thematic Apperception Test) images elicitation: the subject had to comment a series of TAT images [14].
- Neutral text reading: subjects read a text that was supposed not to elicit a strong emotional reaction

Each task lasted approximately from 3 to 5 minutes. Audio signals were recorded with two high quality directional microphones. One microphone was used to record clinician speech and to allow automatic detection of patient speech. The sample frequency was equal to 48 KHz with a 32 bits resolution. The algorithm was applied to obtain average pitch, jitter and pitch SD for each voiced segment detected from the audio file. An intra-subject statistical analysis was performed to investigate changes in the speech features between the two sessions. No comparisons were made between the features extracted from the sub-sessions recordings that are related to different tasks. A Kolmogorov-Smirnov test was applied to verify the normality of the features distributions. A Mann-Whitney U test was adopted for non-gaussian features, otherwise a *t*-test was adopted. Null hypothesis of equality of means or of medians, was accepted if the obtained *p*-value was higher than 0.05.

III. Experimental Results

A. CMU Arctic Database

The pitch estimates obtained from the EGG signal were compared with those obtained from the audio files. The speech corpus considered for the test, contains more than 8000 vowels. The results obtained from the analysis of speech

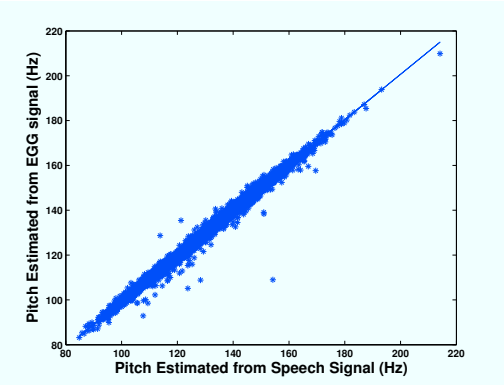

Fig. 1. Pitch from audio signals (*x*-axis) compared with pitch from EGG.

segments are shown in figure 1. The correlation coefficient between the two values is 0.99 and the fitted linear regression model has a slope equal to 1.01 and an intercept equal to -2.9.

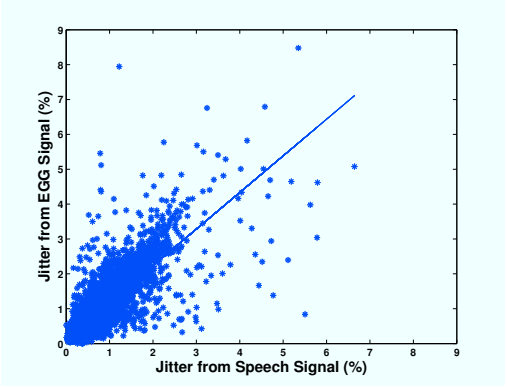

Fig. 2. Jitter from audio signals (x-axis) with respect jitter from EGG.

The results related to jitter estimation are shown in figure 2. The correlation coefficient between jitter estimated with the proposed approach on speech signals and jitter as estimated from EGG was found to be equal to 0.82. The slope is equal to 1.05 and the intercept to 0.13. The correlation coefficient between pitch SD estimated with the proposed approach on speech signals and pitch SD estimated from EGG was found to be equal to 0.93. Results are shown in figure 3. The slope is equal to 0.80 and the intercept is equal to -0.11.

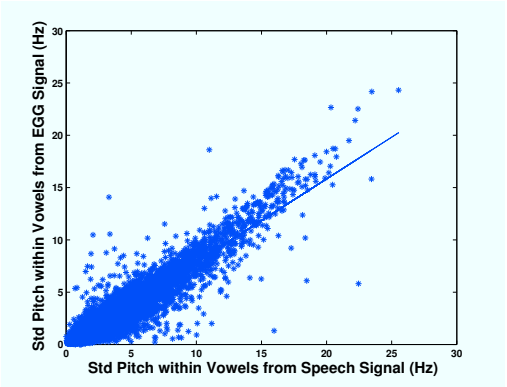

Fig. 3. Pitch standard deviation (pitch SD) from audio signals (x-axis) compared with pitch SD from EGG.

B. Speech data

In this section preliminary results on patients are described. All the subjects were in a different mood state in the second recording session with respect to the first one. In particular in the first recording session subjects A, B and C were scored as hypomanic while subjects D, E and F were scored as depressed. All subjects were scored as euthymic in the second recording session. Pitch values have shown to be normally distributed while both jitter and within voiced segment pitch SD, have been found to have non Gaussian distribution. Results are reported in the included tables. Statistically significant differences between recordings related to the same task are highlighted with a symbol. When both tasks show statistically significant differences, two symbols are used (* or †). The analysis was performed at single subject level and statistically significant differences were highlighted. In particular pitch was found to be higher in hypomanic phase with respect to euthymic state (table I). Moreover pitch was higher in euthymic with respect to depressed phase for subjects D and F, while the opposite trend was observed in subject E. Differences were not found consistently for all the different tasks. In fact for subjects A, C and F these were found only in neutral reading recordings, while in subjects B and D only TAT recordings led to a statistically significant difference. Both jitter and pitch SD (tables II and III) were found to be higher in hypomanic with respect to euthymic in subjects B and C, while no differences in jitter values were found in subject A. Jitter was found to be higher in depressed with respect to euthymic phase in subjects D and F. The opposite behavior was observed in subject E only for TAT recording. As for pitch, the analysis of jitter and pitch SD revealed that the differences were not observed consistently across tasks. In fact jitter was found to be statistically different during TAT recording five times on six, while statistically significant differences during neutral text recordings were observed twice. As regards pitch SD, differences limited only to TAT recordings were observed in subjects B and D, while for subject A a statistically significant difference was highlighted only analyzing the neutral text recording.

TABLE I
MEAN AND SD OF PITCH ESTIMATED FROM VOICED SEGMENTS (HZ).

Id	Session1		Session2	
	Reading	T.A.T.	Reading	T.A.T.
A	Hypomanic		Euthymic	
	131 ± 12*	131 ± 20	119 ± 12*	129 ± 22
B	Hypomanic		Euthymic	
	114 ± 19	133 ± 22*	112 ± 23	105 ± 19*
C	Hypomanic		Euthymic	
	100 ± 12*	102 ± 14	98 ± 9*	102 ± 13
D	Depressed		Euthymic	
	105 ± 7	105 ± 10*	104 ± 7	111 ± 8*
E	Depressed		Euthymic	
	105 ± 8*	$116 \pm 8^\dagger$	100 ± 11*	$99 \pm 10^\dagger$
F	Depressed		Euthymic	
	187 ± 19*	186 ± 31	211 ± 29*	185 ± 14

TABLE II
MEDIAN AND MAD OF JITTER ESTIMATED FROM VOICED SEGMENTS (%).

Id	Session1		Session2	
	Reading	T.A.T.	Reading	T.A.T.
A	Hypomanic		Euthymic	
	0.72 ± 0.42	0.71 ± 0.40	0.78 ± 0.45	0.76 ± 0.44
B	Hypomanic		Euthymic	
	1.08 ± 0.66	1.05 ± 0.62*	1.31 ± 0.89	0.77 ± 0.38*
C	Hypomanic		Euthymic	
	1.08 ± 0.53*	$0.91 \pm 0.55^\dagger$	0.95 ± 0.46*	$0.82 \pm 0.47^\dagger$
D	Depressed		Euthymic	
	0.69 ± 0.31	0.79 ± 0.43*	0.72 ± 0.33	0.71 ± 0.36*
E	Depressed		Euthymic	
	0.87 ± 0.54	0.76 ± 0.50*	0.96 ± 0.56	1.08 ± 0.66*
F	Depressed		Euthymic	
	0.61 ± 0.27*	$0.66 \pm 0.39^\dagger$	0.47 ± 0.52*	$0.49 \pm 0.26^\dagger$

TABLE III
MEDIAN AND MAD OF PITCH SD ESTIMATES (HZ).

ID	Session1		Session2	
	Reading	T.A.T.	Reading	T.A.T.
A	Hypomanic		Euthymic	
	4.01 ± 2.69*	3.86 ± 2.57	3.13 ± 2.00*	3.99 ± 2.80
B	Hypomanic		Euthymic	
	4.81 ± 3.59	5.93 ± 4.20*	5.84 ± 4.72	2.67 ± 1.56*
C	Hypomanic		Euthymic	
	3.71 ± 2.36*	$3.71 \pm 2.43^{\dagger}$	3.43 ± 2.09*	$3.29 \pm 2.15^{\dagger}$
D	Depressed		Euthymic	
	2.61 ± 1.50	3.08 ± 1.82*	2.66 ± 1.51	2.39 ± 1.38*
E	Depressed		Euthymic	
	3.47 ± 2.30	4.22 ± 3.20	3.45 ± 2.39	4.40 ± 3.10
F	Depressed		Euthymic	
	4.90 ± 2.80	4.7 ± 2.92	5.30 ± 3.50	3.93 ± 2.14

IV. DISCUSSION AND CONCLUSION

Results obtained from the speech database, revealed that the proposed approach can be used to estimate pitch values quite robustly. A high correlation value between pitch SD estimated with the proposed approach and the value obtained from the EGG signal was obtained as well. The correlation coefficient between estimated jitter and the benchmark value obtained by EGG signal was lower, even if statistically significant. We have to stress that the proposed approach estimates one pitch value within a time window of 4 pitch periods length. This procedure on the one hand allows achieving robust results with respect to noise, on the other hand results in a systematic jitter underestimation [12]. However, since we are interested in highlighting the differences in glottal features between different mood states, this may not represent a limitation.

Some preliminary results on patients are introduced. All subjects were recorded in different days, and they all were assessed in a different mood state in the different sessions. Given the limited number of subjects, it was not possible to perform a group level analysis, but statistically significant differences at single subject level were investigated. The intra-subject analysis was thus performed to reveal possible differences, from speech samples acquired in the same task category, between the different sessions. Even if the limited number of subjects enrolled limits the generalizability of the results, some observations can be made. The analysis of pitch revealed statistically significant differences between different mood states in all subjects, at least in one of the two tasks considered. This was true also for jitter in all but one subject. Moreover when the change was observed in both tasks, its direction was found to be the same. Looking more in detail at the change sign between mood states and comparing it with the literature results, we believe that it is possible to gain some useful information to get a deeper insight into the pathophysiological meaning of the proposed features. In particular it was shown that the observed changes are not consistent across subjects. In fact pitch was lower in depressed state in two subjects while the opposite behavior was observed for the subject E. The former trend is more frequently reported in literature, but the latter has been observed and associated with coexisting anxiety [6]. This observation confirms the need of improving the characterization of subjects' psychological status and of taking into account anxiety dimension to clarify possible interactions between mood and anxiety levels. Jitter was found to be lower in euthymic state with respect to the other states in all but two cases. In particular in one subject, jitter was found to be equal in hypomanic and euthymic state, and in subject E it was found to be higher in euthymic with respect to depressed state. As for pitch, the latter case is in contradiction with results more frequently reported in literature. Another observation is related to different trends observed in features changes obtained considering different task conditions. In fact, for some subjects a statistically significant difference was seen only for one task condition, as for jitter in subjects D and E. This observation poses some questions concerning the task choice that may interact, at least for some subjects, with the mood state. Overall, these preliminary results seem to indicate the interest in studying glottal related features at single subject level and the need of an improvement in the characterization of each subject psychological status.

REFERENCES

[1] R. Belmaker, "Bipolar disorder," *New England Journal of Medicine*, vol. 351, no. 5, pp. 476–486, 2004.

[2] C. Sobin and H. Sackeim, "Psychomotor symptoms of depression." *The American Journal of Psychiatry*, vol. 154, pp. 14–17, 1997.

[3] D. France, R. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, 2000.

[4] Å. Nilsonne, J. Sundberg, S. Ternstrom, and A. Askenfelt, "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression." *J. Acoustical Society of America*, 1988.

[5] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain and Cognition*, vol. 56, no. 1, pp. 30–35, 2004.

[6] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96 –107, jan. 2008.

[7] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *J. Neurolinguistics*, vol. 20, pp. 50–64, Jan 2007.

[8] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 9, pp. 1530–1540, 2004.

[9] A. Camacho and J. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[10] N. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[11] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *Acoustics, Speech and Signal Processing, IEEE Trans.*, vol. 24, no. 3, pp. 201–212, jun 1976.

[12] 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 Int. Workshop Models and Analysis of Vocal Emissions for Biomedical Applications*, 2009, pp. 29–32.

[13] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis CMU Language Technologies Institute," *Language Technologies Institute, CMU, Pittsburgh PA, Tech Report CMU-LTI-03-177*, 2003.

[14] H. Murray, "Uses of the thematic apperception test." *The American Journal of Psychiatry*, vol. 107, no. 8, pp. 577–581, 1951.