Prosodic Analysis of Speech and the Underlying Mental State

Roi Kliper[1], Shirley Portuguese[2], and Daphna Weinshall[1]

[1]School of Computer Science and Engineering, Hebrew University of Jerusalem, 91904 Jerusalem, Israel
[2]MacLean Psychiatric Hospital, Boston 02478, USA
{kliper@cornell.edu,shirport@walla.com,daphna@cs.huji.ac.il}

Abstract. Speech is a measurable behavior that can be used as a biomarker for various mental states including schizophrenia and depression. In this paper we show that simple temporal domain features, extracted from conversational speech, may highlight alterations in acoustic characteristics that are manifested in changes in speech prosody - these changes may, in turn, indicate an underlying mental condition. We have developed automatic computational tools for the monitoring of pathological mental states - including characterization, detection, and classification. We show that some features strongly correlate with perceptual diagnostic evaluation scales of both schizophrenia and depression, suggesting the contribution of such acoustic speech properties to the perception of an apparent mental condition. We further show that one can use these temporal domain features to correctly classify up to 87.5% and up to 70% of the speakers in a two-way and in a three-way classification tasks respectively.

Key words: Schizophrenia, Machine learning, Mental health, speech prosody, jitter, shimmer

1 Introduction

Psychiatry is a medical discipline in search of objective and clinically applicable assessment and monitoring tools. The acoustic characteristics of speech are a measurable behavior, hence can be used in the assessment and monitoring of disorders such as schizophrenia and depression. This observation has not gone unnoticed in the psychiatric community and previous attempts to quantify this acoustic effect in the psychiatric setting have been made. However, these attempts have been limited, in part by technical and technological limitations. Recent technological advancement has made the recording, storage and analysis of speech an available option for both researchers and practitioners.

The use of acoustic characteristics of speech in the description of pathological voice qualities has been studied in various contexts and with a variety of goals including mental health evaluation [1]. Studies have correlated acoustic features with perceptual qualities [2] and to a lesser extent with physiologic conditions at

the glottis [3]. These studies looked at the use of syntactic structures, richness of vocabulary, time to respond and many other qualities.

Speech prosody is the component of speech that refers to the way words are spoken. It includes the rhythm, stress, and intonation of speech. Prosody may reflect various features of the speaker or the utterance: the emotional state of the speaker; the form of the utterance (statement, question, or command); the presence of irony or sarcasm; emphasis, contrast, and focus; or other elements of language that may not be encoded by grammar or choice of vocabulary. Changes in the acoustic characteristics of speech prosody in the course of mental disorders, notably depression and schizophrenia, are a well documented phenomenon [1, 4, 5, 2], and the evaluation of aspects of speech constitutes, today, a standard part of the mental status examination.

The acoustic changes in schizophrenia patients' speech are currently conceptualized as a component of the negative symptoms [6]. The most accepted scale for negative symptoms is the Scale for the Assessment of Negative Symptoms (SANS) [7]. Negative symptoms are divided into five domains including blunted affect, alogia, asociality, anhedonia, and avolition [8], where speech acoustic changes are especially reflected in two different domains - blunted affect (diminished expression of emotion) and alogia (poverty of speech).

Speech prosody is currently measured by subjective rating scales requiring highly trained staff. Several attempts at using speech cues for the automatic quantification of specific mental effects have been made in the past [9, 4, 10, 2]. These attempts have made an effort to first correlate specific acoustic measures to their perceptual (clinical) counterparts, and second to quantify and asses different aspects of subjects' speech using different acoustic measures [5]. The advantage of automatic quantification of effects apparent in different mental states has been highlighted as early as 1938 [11] and is very well outlined in [12].

In [10] lexical analysis was proposed as a measure of mental deficits; but while some success has been shown, it has been claimed and shown that significant aspects of speech are missed when focusing on the lexical level [13]. In [4] several acoustic features were extracted from both structured speech and semi structured interview of depression subjects. A later study [5] showed high correlations between basic prosody measures (mainly inflection) and clinical ratings of negative symptoms of schizophrenia. In these and other studies results are highly task specific. Lastly, in [2] *Inflection* and *speech rate* were identified as discriminative features between schizophrenia patients and controls.

Our goal to develop automatic computational tools for the evaluation of mental state required two corresponding efforts. First, study the physical signal properties specific to schizophrenia and depression. We chose to focus on temporal domain features (see Section 3.1); these features, while not easy to extract or measure, provide a meaningful interpretation and may be referenced in the context of existing clinical evaluation scales.[1] Second, develop an auto-

[1] While it is possible to use existing automatic systems to produce a high dimensional non-specific description of the voice signal, we focus on a small set of meaningful features for two reasons: (i) These features appear to be ecologically relevant

matic, real-time, reliable and objective assessment of the signal. We adopted a discriminative approach and trained a Support Vector Machine (SVM) classifier over the data (see Section 3.3).

2 Materials and Methods:

Subjects 62 subjects participated in the study, giving written consent approved by the McLean Hospital Institutional review board. The study subjects comprised of three groups, including schizophrenia patients (n=22, 13 male, 9 female), patients with clinical depression (n=20, 9 male, 11 female), and healthy participants (n=20, 10 male, 10 female). The subjects were matched by age ($mean = 39.98, std = 11.37, p = 0.8489$), years of education ($mean = 14.8, std = 2.3, p = 0.063$), and gender ($\chi^2$ test of Independence, $q = 0.86$, $dof = 3$, $p = 0.65$).

Clinically rated symptom measures The subjects completed a clinical interview which included Semi structured Clinical Interview for DSM-IV (SCID IV) [14] , Positive and negative Syndrome Scale for Schizophrenia (PANSS) [15], Scale for the Assessment of Negative Symptoms (SANS) [7], and the Montgomery and Absberg Depression Rating Scale (MADRS) [16] as well as the Hamilton Depression Scale (Ham-D) [17].

Acoustic Recordings The recordings were made by a headset without sound isolation or calibration. To prepare the recordings for acoustic analysis, the audio tapes were digitized at a 44.1 kHz sampling rate. Acoustic analysis was conducted using MATLAB [18] (details are given in Section 2.1). Average length of a clinical interview was: schizophrenia - 57m 13s, depression - 30m 31s, healthy - 48m 46s. Silence was automatically removed at the beginning and end of each recording, while the remaining data was normalized to have 0 mean and variance 1, thus avoiding effects caused by the constellation of the headset. To enable efficient handling of the data each interview was divided into 2 minutes segments, which were subsequently analyzed independently. All results from a single person's 2 minutes segments were later used together for classification.

2.1 Speech Prosody and Feature Extraction

Two minutes segments of interview were used to measure nine diagnostic features. In this paper we focus on a very small and simple set of features extracted

and correspond with psychiatrists' intuition about the characteristic features of the speech of Schizophrenia patients. (ii) Our application domain suffers from the problem of small sample, which necessitates the use of low dimensional representations to enable effective learning; this is accomplished by choosing a small set of relevant features. The alternative, which is to use a high dimensional representation followed by dimensionality reduction (like PCA), typically leads to the unfortunate outcome that the final result is hard to interpret in terms of the underlying features.

in the temporal domain. This choice was motivated by the fact that temporal domain features are, in general, easier to relate to perceptual properties of speech and thus provide better infrastructure for further use in the psychiatric community.

Alterations of the speech signal in abnormal conditions can occur at different time-scale levels, including the *macro-scale* level (above 1 sec) which refers to variables such as speaking rate, the *meso-scale* (25 ms to 1 sec), in which variables like pitch and its statistics are measured, and finally, the *micro-scale* (10 ms or less) level in which cycle to cycle measures are taken (this level appears to contribute to the naturalness of the speech sound). While *macro* and *meso* scales are influenced by voluntary aspects of speech, the *micro-scale* is involuntary in nature and, thus, can better serve as a reliable biomarker. All scales contribute to the prosodic structure of the speech signal, and thus using them as an ensemble may provide insight into the possible role of prosody in the characterization of pathological mental states. The focus on prosodic features follows reports showing the relevance of meso-scale and micro-scale levels to the tasks of mental evaluation and emotion detection [1, 2, 3].

Macro scale measures: Mean utterance duration, Mean gap duration, Mean spoken ratio An utterance is any segment identified as speech that exceeds 0.5 seconds. A "gap" is any segment of recording with no subjects' speech. *Spoken ratio* is calculated as the ratio between the total volume of speech occupied by the speaker, that is the sum of the length of all utterances divided by the total conversation length.

Meso scale measures: Pitch Range, Pitch standard deviation, Power standard deviation *Pitch range* was calculated as the difference between maximum estimated pitch and minimum estimated pitch normalized by the mean pitch over the entire 2 minutes segments. No significant between-group differences were observed for mean pitch, which was (Males: 112.3Hz, 18.99Hz; Females: 176.39Hz, 18.53Hz) ($F = 0.18, df = 2, 56, p = 0.83$) nor for interaction with gender ($F = 0.75, df = 2, 56, p = 0.47$). Between gender differences were strong as expected ($F = 182.9, df = 1, 56, p << 0.01$). Standard deviation (STD) of pitch was calculated for each utterance and was then averaged for each speaker. STD of pitch was again normalized by the mean pitch in each utterance. Mean active power and its variance were measured in decibels (dB) in reference to a calibration level and were 64 dB, with a standard deviation of 8.6 dB. Standard deviation of power within an utterance was measured, normalized by the mean power in the entire segment in order to avoid effects caused by noise in the location of the microphone.

Micro scale measures: Mean waveform correlation, Mean jitter, Mean shimmer The mean of all correlation coefficients evaluated for every pair of consecutive periods was used as the acoustic measure termed *Mean Waveform Correlation (MWC)*. It indicates the overall similarity between the cycles of the time signal. When applied to pitched segments it measures the level at which the speaker sustained its constant pitch.

Following [1] jitter and shimmer were calculated as the period perturbation quotient PPQ and the energy perturbation quotient EPQ respectively, where the locality parameter was chosen to be 5. Perturbation Quotient (PQ) measures the local deviation from stationarity of a given measure and is defined in (1). It measures the ratio of deviation of a given measure in a local neighborhood defined by the locality parameter K. Put in simple terms, the jitter measures the stability of the period in a 5-local cycles environment and the shimmer measures the stability of the energy in a given 5-local cycles environment.

$$PQ = \frac{100\%}{N-K} \sum_{\nu=\frac{K-1}{2}}^{N-\frac{K-1}{2}-1} \left| \frac{u(v) - \frac{1}{K}\sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} u\left(v+k\right)}{\frac{1}{K}\sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} u\left(v+k\right)} \right| \tag{1}$$

We chose to use the energy shimmer since it is expected to be considerably less susceptible to noise than the amplitude shimmer often used in acoustic analysis.

Discussion: We have previously observed that both macro-scale and meso-scale measures seem to incorporate a larger variability component due to the specific task: between task-variability S_b as compared to the within speaking-task variability (S_w). This task dependency is significantly reduced for micro-scale measures. This observation does not disqualify the larger scale measures from being useful in a classification task (as the task is known in advance); however, it highlights micro-scale features as candidates to be used by a robust general-purpose classifier.

2.2 Classification and Statistical Analysis

We used the extracted acoustic features and a basic linear classifier [19] to classify the different conditions: Healthy (HL), Schizophrenia (SZ) and Depression (DP) in a two-way classification task. For each classification scenario (e.g. HL vs. SZ), one subject was left out for testing and the rest were used to train a classifier. In a single classification setting all 2 minutes segments of a given speaker were left out and later used for testing. The final decision was taken using a majority vote over all left out segments.

To check the statistical significance of the results over each of the individual features, we used 1-way ANOVA when the distribution was roughly normal, otherwise we used the nonparametric version of the 1-way ANOVA called the Kruskal-Wallis test. It is actually a more general test, in that it is comparing distributions rather than medians. Checking the statistical significance of the effects brings up the problem of multiplicity (multiple comparisons). We consider the problem of testing simultaneously 9 null hypotheses where within each 3 pair comparisons are nested. We therefore used a sequential Bonferroni type procedure, which is very conservative and assures the statistical soundness of the results. Specifically, the results were first tested for significance of main effect

using a Bonferroni correction, and later multiple comparison was done using a second Bonferroni correction[2].

3 Results

In the following section we describe first an analysis of the individual features that were extracted as indicated above. We start by describing the distributions of the different features according to the different groups in Section 3.1. In Section 3.2 we describe the correlations between these features and standard clinical ratings, while in Section 3.3 we describe our efforts to train a an SVM classifier to predict the speaker's condition.

3.1 Isolated Features - between group analysis

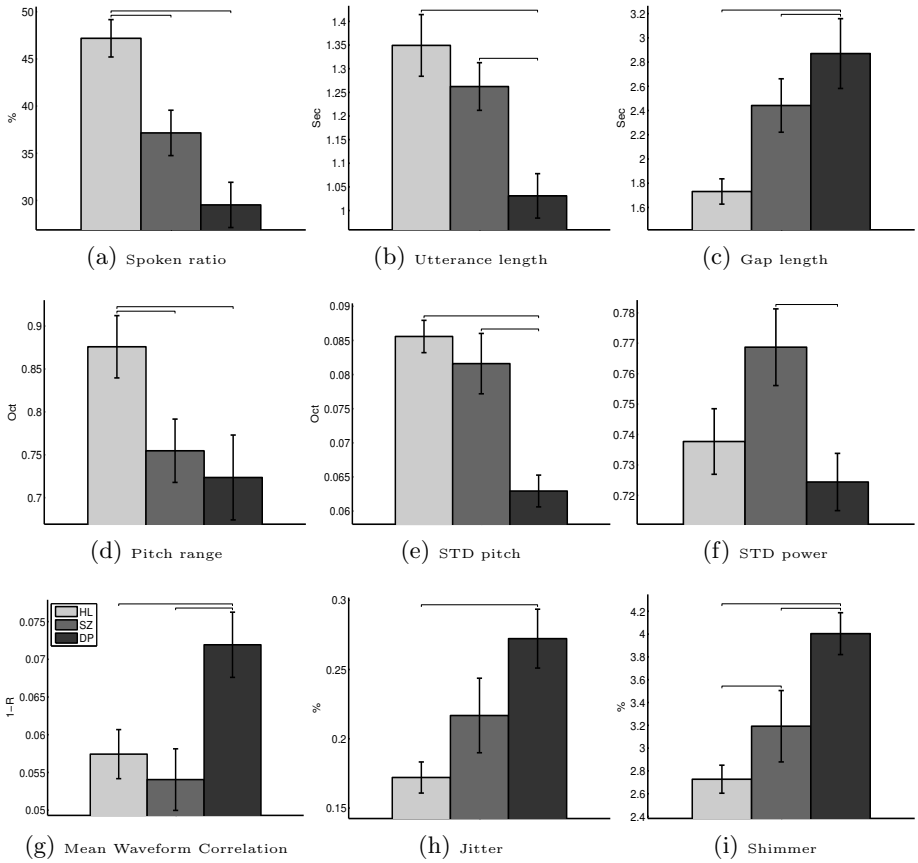
Fig. 1 shows the mean and standard error of isolated features extracted from semi structured interview. Statistically significant deviations between any two groups are indicated with a horizontal bar. The analysis was performed while taking into consideration the issue of multiple comparisons as explained in Section 2.2, and is thus very conservative in nature.

Spoken ratio As seen in Fig. 1a, the ratio of spoken volume for healthy subjects (47.19%,1.98%) is larger than that of Schizophrenia subjects (37.16%,2.4%) and Depressed subjects (29.52%, 2.4%). The difference between the groups is indeed significant ($\chi^2 = 21.63; df = 2, 59; p < 0.001$) and possibly reflects the subject's initiative and willingness to engage in conversation.

Utterance Duration Healthy subjects appear to speak in longer utterances ($\sim$1.35s, 0.07s) as compared to Schizophrenia subjects (1.26s, 0.05s) and depressed subjects (1.03s,0.05s). Significant ($\chi^2 = 16.06; df = 2, 59; p < 0.001$) differences were observed between the depressed group and both the group of healthy subjects and schizophrenia subjects. Only a trend was observed between normal controls and schizophrenia subjects. In depressed recordings an average utterance length that exceeds 3 seconds (as averaged over a two minutes segment) never occurred, which is reflective of the reported difficulty of engaging in conversation. These results tend to agree with previous reports [20].

Gap Duration Normal subjects tend to pause less and for less time (1.73s, 0.1s) as compared to schizophrenia subjects (2.44s,0.22s) and depressed subjects (2.87s,0.29s) ($\chi^2 = 14.63$; $df = 2, 59$; $p < 0.001$). Also, the pauses of healthy subjects are of more regular pattern as reflected by the small error bar.

[2] Note that with only three treatment groups, it's overly conservative to adjust the alpha levels with a Bonferroni method as with only 3 treatment groups, there is little risk in an increasing Type I error rate.



(a) Spoken ratio (b) Utterance length (c) Gap length

(d) Pitch range (e) STD pitch (f) STD power

(g) Mean Waveform Correlation (h) Jitter (i) Shimmer

Fig. 1: **Individual features:** 9 features are analyzed showing mean and standard error (STE) bar for each feature in each group of patients. Significant differences are indicated with a horizontal bar. Significance was established using a Kruskal-Wallis procedure and a Bonferonni correction for multiple comparisons.

Pitch Range Healthy subjects evidently displayed larger pitch range of (~ 0.88, ~ 0.04) (mean, ste) of an octave, which is in agreement with reported literature [9, 21, 5]. Pitch range is significantly reduced for schizophrenia (0.75, $\sim$0.04) with an even lower pitch range of (0.67,$\sim$0.05) for depressed patients ($\chi^2 = 7.49$; $df = 2, 59$; $p < 0.05$).

Standard Deviation of Pitch Our measure of standard deviation of pitch within an utterance represents a temporally local perception of inflection. healthy subjects display a wider dynamics of pitch (0.0992,0.0024) within an utterance as compared to schizophrenia subjects (0.0924, 0.0044) and depressed subjects are significantly different (0.0629,0.0023) ($\chi^2 = 21.28$; $df = 2, 59$; $p < 0.001$).

Power standard deviation Differences in power standard deviation were evident between Schizophrenia subjects ($\sim$0.77, $\sim$0.01) and Healthy subjects (0.74,$\sim$0.01) on the one hand, and Depressed subjects (0.7244,0.01) on the other hand ($\chi^2 = 6; df = 2, 59; p < 0.05$). Only differences between schizophrenia and depression remained significant after correction for multiple comparisons.

Mean Waveform Correlation The results show a significant deviation of the depressed subjects ($\chi^2 = 10.42; df = 2, 59; p < 0.01$).

Mean Jitter We see significant differences between healthy subjects (0.1722, $\sim$ 0.01) and both schizophrenia subjects ($\sim$0.22, 0.01) and depressed subjects ($\sim$0.27, 0.0212). The way jitter was calculated puts a focus on the physiological ability to maintain a constant period, and suggests a deficiency in this ability in both schizophrenia and depression subjects.

Mean Shimmer Healthy subjects displayed the lowest shimmer ($2.73\%, 0.12\%$) whereas depressed subjects displayed an elevated shimmer ($4\%, 0.18\%$) with an intermediate level ($3.22\%, 0.12\%$) for schizophrenia subjects. Again these results may suggest some problem in spontaneous control of the glottal production mechanism. All post-hoc between-group comparisons were significant ($\chi^2 = 26.41; df = 2, 59; p << 0.01$).

3.2 Correlation

In order to compensate for excessive skew in the clinical measures we followed [2] and employed non-parametric statistics (Spearman's ρ correlation coefficient). Rank order correlations (Spearman) were computed between the acoustic and clinical based symptoms of the subjects (this data is omitted). More interestingly, we correlated the acoustic measures with the diagnostic rating as seen in Table 1. Here the correlation scores were only calculated within the relevant group, that is, Schizophrenia clinical ratings were correlated with acoustic measures of subjects diagnosed with schizophrenia, while depression clinical ratings were correlated with acoustic measures of depressed subjects only.

Some findings in Table 1 are worth special mention. Spoken ratio was defined to agree with the description of Alogia as a reduction in quantity of speech; we find it reassuring that it is highly correlated with the SANS-alogia clinical rating ($0.64, p << 0.01$). Contrary to reported results in [2], high correlations between STD of pitch and spoken ratio were observed.

3.3 Classification

The linear support vector machines (SVM) classifier [22] was employed to train discriminative models using the extracted measures. The task consists of either binary classification, where a model was trained to discriminate between two distinct mental states, or multi-class classification, where a set of models was trained to identify the mental state of a specific speaker in a *1-vs.-all* approach.

	Schizophrenia Scales				Depression Scales	
		SANS				
	PANSS	(total)	(affect)	(alogia)	MADRAS	HAM-D
1. Spoken Ratio	-0.11	-0.5^{**}	-0.58^{**}	-0.64^{**}	-0.11^{*}	0.11
2. Utterance Duration	-0.19	-0.44^{**}	-0.55^{**}	-0.49^{**}	-0.27^{*}	0.05
3. Gap Duration	0.09	0.45^{**}	0.45^{**}	0.54^{**}	-0.01^{*}	-0.14
4. Pitch Range	-0.16	0.04	0.13	0	-0.35^{*}	-0.33^{*}
5. STD Pitch	0	-0.07	-0.17	-0.17	-0.18	-0.15
6. STD Power	-0.14	-0.27	-0.39	-0.31^{*}	-0.01	0.1
7. $1 - MWC$	0.03	0.24	0.4	0.32^{*}	0.19	0.19
8. Jitter	0.34	0.38^{*}	0.21	0.45^{*}	-0.1	-0.3
9. Shimmer	0.41^{*}	0.4^{*}	0.18	0.31^{*}	-0.01	0.2

Table 1: **Acoustic Features- Psychiatric Scales Correlations**: * indicates $p < 0.05$, ** indicates $p << 0.01$.

Our classifier obtained the following pair-wise classification success rates (chance at 50%): control vs. Schizophrenia - 76.19%, control vs. depression - 87.5%, and Schizophrenia vs depression - 71.43%. Multi-class classification success rate was at 69.77 (chance at 33.3%).

4 Summary and Discussion

Speech acoustics is a measurable behavior that could be utilized as a biomarker in the clinical setting. The change in the acoustics of speech is not the only aspect of speech that changes in the course of various disorders [21], but these changes are a well documented phenomenon in both schizophrenia and depression. In both disorders speech acoustics often changes over time.

Our study was motivated by the desire to contribute to the search for a possible biomarker for schizophrenia and major depressive disorder. Clearly the development of reliable, objective, low-priced, and readily applicable assessment tools would enhance the accuracy of the clinical evaluation for diagnosis and monitoring. We focused on a relatively simple set of features extracted from the speech signal in the temporal domain. We divided the set of features into three groups of features, according to the time scale required for their extraction. We showed that while macro-scale features correlate with distinct components of the SANS rating scale, meso-scale features show poor correlations. Micro-scale features showed the highest promise as diagnostic measures both in terms of reliability and validity. Our findings that the acoustic features can separate schizophrenia from depression subjects, without reference to the content of the speech, provides converging evidence for the promise of this approach.

Acknowledgements: This work was supported in part by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI), and the Gatsby Charitable Foundations.

References

1. Michaelis, D., Fröhlich, M., Strube, H.: Selection and combination of acoustic features for the description of pathologic voices. The Journal of the Acoustical Society of America **103** (1998) 1628
2. Cohen, A., Alpert, M., Nienow, T., Dinzeo, T., Docherty, N.: Computerized measurement of negative symptoms in schizophrenia. Journal of psychiatric research **42** (2008) 827–836
3. Moore, E., Clements, M., Peifer, J., Weisser, L.: Critical analysis of the impact of glottal features in the classification of clinical depression in speech. Biomedical Engineering, IEEE Transactions on **55** (2008) 96–107
4. Alpert, M., Pouget, E., Silva, R.: Reflections of depression in acoustic measures of the patient's speech. Journal of affective disorders **66** (2001) 59–69
5. Alpert, M., Shaw, R., Pouget, E., Lim, K.: A comparison of clinical ratings with vocal acoustic measures of flat affect and alogia. Journal of psychiatric research **36** (2002) 347–353
6. Association, A.P., on DSM-IV., A.P.A.T.F.: Diagnostic and statistical manual of mental disorders: DSM-IV-TR. American Psychiatric Publishing, Inc. (2000)
7. Andreasen, N.: Scale for the assessment of negative symptoms (sans). British Journal of Psychiatry (1989)
8. Andreasen, N.: Negative symptoms in schizophrenia: definition and reliability. Archives of General Psychiatry **39** (1982) 784
9. Andreasen, N., Alpert, M., Martz, M.: Acoustic analysis: an objective measure of affective flattening. Archives of General Psychiatry **38** (1981) 281
10. Pennebaker, J., Mehl, M., Niederhoffer, K.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology **54** (2003) 547–577
11. Newman, S., Mather, V.: Analysis of spoken language of patients with affective disorders. American Journal of Psychiatry **94** (1938) 913
12. IsHak, W.: Outcome measurement in psychiatry: A critical review. American Psychiatric Pub (2002)
13. Kring, A., Bachorowski, J.: Emotions and psychopathology. Cognition & Emotion **13** (1999) 575–599
14. DSM, I.: American psychiatric association. Diagnostic and Statistical Manual of Mental Disorders (1994)
15. Kay, S., Flszbein, A., Opfer, L.: The positive and negative syndrome scale (panss) for schizophrenia. Schizophrenia bulletin **13** (1987) 261
16. Montgomery, S., Asberg, M.: A new depression scale designed to be sensitive to change. The British Journal of Psychiatry **134** (1979) 382
17. Hamilton, M.: A rating scale for depression. Journal of neurology, neurosurgery, and psychiatry **23** (1960) 56
18. MATLAB: version 7.11.0 (R2010b). The MathWorks Inc., Natick, Massachusetts (2010)
19. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research **9** (2008) 1871–1874
20. Alpert, M., Kotsaftis, A., Pouget, E.: Speech fluency and schizophrenic negative signs. Schizophrenia Bulletin **23** (1997) 171–177
21. Alpert, M., Rosenberg, S., Pouget, E., Shaw, R.: Prosody and lexical accuracy in flat affect schizophrenia. Psychiatry Research **97** (2000) 107–118
22. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874