Evidence for Depression and Schizophrenia in Speech Prosody*

Roi Kliper[1], Yonatan Vaizman[1], Daphna Weinshall[1] and Shirley Portuguese[2]
[1]The Interdisciplinary Center for Neural Computation and Computer Science Department, The Hebrew University of Jerusalem, Israel.
[2]Maclean Psychiatric Hospital, Boston, USA.

Abstract

We developed automatic computational tools for the monitoring of pathological mental states – including characterization, detection, and classification. We show that simple temporal domain features of speech may be used to correctly classify up to 80% of the speakers in a two-way classification task. We further show that some features strongly correlate with certain diagnostic evaluation scales, suggesting the contribution of such acoustic speech properties to the perception of an apparent mental condition.

Key words: Speech analysis, Schizophrenia, Depression.

Introduction

Changes in the acoustic characteristics of speech prosody in the course of different mental disorders, notably depression and schizophrenia, are a well-documented phenomenon (Alpert, Pouget, & Silva, 2001; Cohen, Alpert, Nienow, Dinzeo, & Docherty, 2008; Michaelis, Fr hlich, & Strube, 1998). The evaluation of speech constitutes a standard part of the repertoire of a mental status examination where standard scales include variables such as Alogia, Affective Flattening and Apparent sadness. Psychiatrists who use these scales, report prosody as an important diagnostic tool in their daily work. However, prosody is currently measured by subjective rating scales acquired by highly trained staff, and as a result these clinical ratings tend to be skewed and restricted in range in comparison to computerized acoustic measures, which may lead to rating bias in the global clinical impression.

In an effort to develop automatic computational tools for the monitoring of these pathological mental states, we followed two parallel and tightly connected research tracks: (1) Study the physical signal properties that highlight the examined mental states, (2) Develop an automatic, real-time, reliable and objective assessment tool for diagnostic screening, illness progression monitoring, and response to treatment assessment.

Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics 2010, 25-27 August 2010, Athens, Greece.

Mental Illness characterization: Data, Feature Extraction

Our data consisted of speech recordings from schizophrenic, depressed and control subjects performing three different tasks: reading a standardized list of words, reading an emotionally neutral passage, and participating in a semi structured clinical interview performed by a clinician. These recordings were used separately to extract various acoustic features that characterize the speech signal.

We focused here on features extracted in the temporal domain. Though spectral features have been proven beneficial in various speech analysis tasks, we leave the examination of these features for future research. Alterations of the speech signal in abnormal conditions can occur at different time-scale levels, Including the meso-scale (25 ms to 1 second) and the micro-scale (10 ms or less) levels. By focusing on these scales, we follow recent reports from the closely related area of speech emotion detection that have shown the relevance of meso-scale and micro-scale level to the task of emotion detection. (Cowie & Cornelius, 2003; Rong, Li, & Chen, 2009). Specifically, for each speaking task we used an energy based Utterance-Silence segmentation from which we extracted utterance and gap statistics (e.g. mean utterance duration). We then used the YIN pitch detection algorithm (de Cheveign & Kawahara, 2002) to detect pitched segments (among uttered segments) and to further segment the Uttered segments into Voiced-Unvoiced segments, followed by the extraction of power and pitch related statistics (e.g. Inflection, Emphasis). We, then, used pitched segments to refer to the finer resolution of 'cycle to cycle' variation and extract micro-scale statistics (e.g. Jitter, Shimmer).

Some researches have claimed that speech measurements have arguable validity and poor reliability, pointing in part to the large within subject variability (Rong et al., 2009). This issue is indeed crucial in all applications of speech analysis and poses a major problem for the development of reliable classification tools. Ideally, for the purpose of classification, we would have liked to employ measures which are (speech) task-independent on the one hand and (mental) condition-dependent on the other hand. We observed that both large-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): $S_b/S_w > 2.8$. This task dependency is significantly reduced for micro-scale measures where one sees the ratio $S_b/S_w < 1.8$ and generally around 1. 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 the possibility that micro-scale features may serve as better candidates for a

robust general-purpose classifier being relatively task invariant. To check the reliability of these measures for classification of the different conditions over general speaking tasks, we examined the variance ratio of the between condition vs. within condition. The higher this ratio is, the more discriminative a measure is for classification. While, in general, we did not get high ratios (due to the general signal to noise ratio), the highest ratios where obtained for micro-scale features, notably shimmer, and mean-waveform-correlation. Features that display such robustness (both task invariance and condition dependence) may serve as efficient indicators of a mental illness.

Results: Classification and Correlations:

We used the extracted acoustic features and a basic linear classifier (Chang & Lin, 2001) to classify the different conditions: Normal (NL), Schizophrenia (SZ) and Depression (DP) in a two-way classification task. For each classification scenario (e.g. NL vs. SZ) and all relevant recordings of a given task (e.g. List of words), one subject was left out for testing and the rest were used to train a classifier. Table 1 contains the average success rate over all leave-one-out test realizations. To account for small training sample size PCA was used to reduce dimensionality of the data-set. As can be seen in the table SZ patients are discriminated nicely from NL with classification rates ranging from 73.81-80.95%. Similar success is achieved for a two-way classification of DP vs. NL. The task of discrimination DP from SZ is reported to be harder but succeeds in two out of three speaking tasks.

Table 1:Classification Results: List of Words (n(c)=20,n(s)=22, n(d)=20), Passage (n(c)=11,n(s)=10, n(d)=15),Interview (n(c)=33 n(s)=31, n(d)=29)

	NL vs. SZ	NL vs. DP	SZ vs. DP
List of words	0.7381	0.70	0.7143
Passage	0.8095	0.69	0.5200
Interview	0.75	0.8710	0.7667

In addition to the strict labels of mental condition, every subject was clinically evaluated by a psychiatrist using standard psychiatric evaluation scales, for instance: Scale for the Assessment of Negative Symptoms (SNAS) for SZ and Hamilton Rating Scale for Depression (HAM-D) for DP. We calculated the correlation coefficients (cc) between the acoustic features extracted from the interview recordings, and the psychiatric evaluations, for the SZ and DP subject groups. Selected results are shown in Table 2.

For SZ, high ratings of negative symptoms correlated with longer gaps and shorter (but denser) utterances and with lower emphasis measures.

For DP, higher depression severity correlated with longer utterances and with lower pitch variability, both in the meso-scale (inflection) and in the micro-scale (jitter).

These correlations are a good indication that the extracted acoustic features capture elements of the speech signal which are perceived by the listener as deviating from normality, and thus may contribute to a better identification and characterization of the underlying mental condition.

Table 2:Correlation Results

Schizophrenia Group	
Acoustic feature	**SANS Total (cc, p)**
Utterance duration	(-0.5953,0.0004)
Gap duration	(0.4223,0.0180)
Spoken ratio	(-0.5134,0.0031)
Fragmented speech	(-0.4831, 0.0059)
Emphasis	(-0.4782,0.0065)
Depression Group	
Acoustic feature	**HAM-D**
Utterance duration	(0.4661, 0.0164)
Inflection	(-0.5007, 0.0092)
Jitter	(-0.4555, 0.0194)

References

Alpert, M., Pouget, E., & Silva, R. (2001). Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders* **66,** 59-69.

Chang, C., & Lin, C. (2001). LIBSVM: a library for support vector machines: Citeseer.

Cohen, A., Alpert, M., Nienow, T., Dinzeo, T., & Docherty, N. (2008). Computerized measurement of negative symptoms in schizophrenia. *Journal of psychiatric research* **42,** 827-836.

Cowie, R., & Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication* **40,** 5-32.

de Cheveign , A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* **111,** 1917.

Michaelis, D., Frohlich, M., & Strube, H. (1998). Selection and combination of acoustic features for the description of pathologic voices. *The Journal of the Acoustical Society of America* **103,** 1628.

Rong, J., Li, G., & Chen, Y. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management* **45,** 315-328

* This study was supported by the European Union under DIRAC integrated project IST-027787