The Raymond and
Beverly Sackler Faculty
of Exact Sciences
Tel Aviv University

Tel Aviv University
Raymond and Beverly Sackler Faculty of Exact Sciences
The Blavatnik School of Computer Science

The Contribution of Prosody to Machine Classification of Schizophrenia

by

Tomer Ben Moshe

Under the supervision of
Dr. Kfir Bar &
Prof. Nachum Dershowitz

Thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science

2022

Abstract

We show how acoustic prosodic features, such as pitch and gaps, can be used computationally for detecting symptoms of schizophrenia from a single spoken response. We compare the individual contributions of speech and text modalities to the determination whether the speaker has schizophrenia. Our classification results clearly show that the prosodic features capture more information than the linguistic ones. We find that, when combined with those prosodic features, linguistic features improve classification only slightly.

Acknowledgements

First and foremost, I would like to express my special thanks and gratitude to my advisors Dr. Kfir Bar and Prof. Nachum Dershowitz for their support and guidance and many insightful conversations during the development of the ideas in this thesis.

Their patience, motivation, and immense knowledge had made me enjoy and learn a lot during this research. I could not have imagined having better advisors and mentors for my M.Sc. study.

Special thanks to the doctors and others at the Psychology Department in Beer Yaakov-Ness Ziona Mental Health Center for their contribution in making this research possible, especially Dr. Ido Ziv for his knowledge sharing and ideas.

Ethics Statement

The institutional review board of the College of Management Academic Studies of Rishon LeZion, as well as of the Beer Yaakov–Ness Ziona Mental Health Center, approved the experiments herein, and informed consent was obtained for all subjects.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Schizophrenia is an acute mental disorder characterized by delusions, hallucinations, and thought disorders. Thought disorders are disturbances in the normal way of thinking, typically presented as various language impairments, such as disorganized speech, which is related to abnormal semantic associations between words [Aloia et al., 1998]. These include the following: (1) poverty of speech; (2) pressure of speech, fast, loud and hard-to-follow responses; (3) "word salad", random-word selection at times; (4) derailment, moving from one topic to another during a conversation; and (5) tangentiality, providing an irrelevant response, never reaching the answer to a question. Andreasen [1979] supplies some statistics for symptoms of thought disorder, with the most common being derailment, loss of goal, poverty of content, and tangentiality.

Diagnosing thought disorders is performed by clinicians and mental-health professionals, typically by means of a conversation. This is an arduous and subjective process. Mental-health professionals are on constant lookout for objective computational assessment tools that can help identify whether a person is showing signs of thought disorders.

There have been previous attempts at developing computational tools for analyzing language with the goal of detecting symptoms of mental-health disorders; we describe some of those works in the following section. Generally speaking, speech and text are the two modalities of a human language that can be processed and analyzed for the diagnosis of mental-health disorders. For this purpose, processing speech is typically done for the purpose of modeling the prosody by extracting features related to intonation, stress and rhythm. One of the most popular prosodic symptoms is flattened intonation, or aprosody, which is interpreted as inability of a person to properly convey emotions through speech. This is a negative symptom of schizophrenia. Another negative symptom that is associated with speech is alogia, or poverty of speech, presented as very minimal speech. Metaphorically, it has been claimed [Cherry, 1964, Spoerri, 1966] that patients with schizophrenia sometimes sound like a person talking on the phone, referring to the low-quality aspect of the voice, sometimes referred to as a "creaky" voice. Cohen et al. [2013] associate acoustic-based analysis of speech, generally speaking, with clinically-rated negative symptoms, while associations with positive symptoms have been found to be inconsistent.

Prosody may reflect elements of language that are not encoded by grammar or by choice of vocabulary. On the other hand, the transcription of speech is typically processed for capturing linguistic and semantic characteristics of the conversation.

In this work, we study the salience of prosodic (from speech) and linguistic (from text) features, for the classification task of automatically detecting whether a given utterance was generated by someone who is diagnosed with schizophrenia, or by control. To do that, we measure the contribution of each set of features once when used individually for classification, and again when both sets are combined together.

Our dataset comprises transcribed interviews, collected from native Hebrew-speaking inpatients officially diagnosed with schizophrenia at a mental health center in Israel, and from a demographically balanced control group. The prosodic features that we use are based on pitch, which we extract using an audio processor. The linguistic features are extracted from the transcription of the audio files, and are designed to capture symptoms such as derailment and incoherence, following a previous work [Bar et al., 2019] that has shown the efficacy of such features when used in a similar classification task.

Prosodic features have been computationally examined previously and were shown to be effective for detecting schizophrenia—for example by Kliper et al. [2010, 2015] for English speech. For Chinese, Huang et al. [2022] combined prosodic features with linguistic features for assessing the severity of thought disorders in examined schizophrenia patients.

However, none of these works compare the individual contributions to classification of each of the modalities when used in combination.

Our contribution is twofold: (1) We show how acoustic prosodic features can be used for detecting symptoms of schizophrenia from only a single spoken response, given in Hebrew; and (2) we measure the individual contribution of both speech and text modalities to the task of detecting whether the person who generated a given utterance has schizophrenia or not. Our classification results clearly show that the acoustic prosodic features capture more information than the linguistic ones. When combined with those prosodic features, linguistic features improve classification only slightly.

Chapter 2

Related Work

The extensive literature about language characteristics and schizophrenia is examined in [Covington et al., 2005]. The authors distinguish between two types of language impairment among patients with schizophrenia: thought disorder—defined as disturbances in the normal way of thinking, and schizophasia—comprising various dysphasia-like impairments such as clanging, neologism, and unintelligible speech. They also claim that patients with thought disorders produce and perceive sounds in an abnormal way, manifesting as flat intonation or unusual voice quality. Hoekert et al. [2007] conducted a meta-analysis of 17 studies between 1980 and 2007. They found that prosodic expression of emotions is significantly impaired with schizophrenia. Martínez-Sánchez et al. [2015] compared the speech of 45 medicated schizophrenia patients and 35 healthy controls, all native Spanish speakers from Spain. The results revealed that patients paused more, talked more slowly, and showed less variability in speech and fewer variations in syllable timing. [Alpert et al., 2000] examined whether "flat affect", defined as emotionless speech, which is one of the symptoms of schizophrenia, indicates an emotional deficiency or whether this is only a communication issue. They did not find evidence for impairment in any other aspect of emotion expression besides prosody.

There is a large body of work that studies the efficacy of computational approaches for diagnosis of mental-health disorders. We continue by listing some related work that use computational tools to process acoustic speech signals for diagnosis of mental-health disorders, followed by works that use natural-language-processing (NLP) tools for analyzing speech transcriptions for the same purpose.

In a systematic review [Low et al., 2020] that analyzes 127 studies, the authors conclude that speech processing technologies could aid mental-health assessment; however, they mention several caveats that need to be addressed, especially the need for comprehensive transdiagnostic and longitudinal studies. Given the diverse types of datasets, feature extraction procedures, computational methodologies, and evaluation criteria, they provide guidelines for both data acquisition and building machine-learning models for diagnosis of mental-health disorders.

Kliper et al. [2015] trained a support vector machine (SVM) classifier that gained about 76% accuracy in a binary classification task of identifying people with schizophrenia and controls, using acoustic features.

The study population comprised 62 English-speaking participants, divided into three groups: patients with schizophrenia, patients with clinical depression, and healthy controls. In a three-way classification task over the three groups, their classifier achieved about 69% accuracy. Every participant was interviewed and recorded by a mental health professional. Each recording was divided into segments of two minutes each, which were subsequently analyzed independently. Each recording was represented by nine acoustic features based on pitch and power, which were automatically extracted using tools similar to those that we use in this work.

Dickey et al. [2012] study prosodic abnormalities in patients with schizoid personality disorder (SPD). Their experimental results showed that SPD patients speak more slowly, with more frequent pauses, and exhibited less pitch variability than control participants.

A new algorithm to detect schizophrenia was proposed by He et al. [2021] based on a classifier that uses three new acoustic prosodic features. On a dataset comprised of 28 schizophrenia patients and 28 healthy controls, they measured classification accuracy between 89.3% and 94.6%.

Agurto et al. [2020] predict psychosis in youth using various acoustic prosodic features, such as pitch-related and Mel-frequency cepstral coefficients (MFCC). They analyzed the recorded speech of 34 young patients who were diagnosed to be at high risk of developing clinical psychosis. Among other things that they showed, they trained a classifier that can predict the development of psychosis with 90% accuracy, outperforming classification using clinical variables only.

There has been an increasing number of works that computationally process speech transcriptions for detecting symptoms of schizophrenia. Specifically, measuring derailment and tangentiality has been addressed several times. For example, Elvevåg et al. [2007] analyzed transcribed interviews of inpatients with schizophrenia by calculating the semantic similarity between the response given the participants and the question that was asked by the interviewer. For similarity they used cosine similarity over the latent semantic analysis (LSA) vectors [Deerwester et al., 1990] calculated for each word, and summed across a sequence of words. Similarly, Bedi et al. [2015] use cosine similarity between pairs of consecutive sentences, each represented by the element-wise average vector of the individual words' LSA vectors, to measure coherence. Using this score they automatically predicted transition to psychosis with 100% accuracy. Iter et al. [2018] showed that removing some functional words from the transcriptions improves the efficiency of using cosine similarity over LSA vectors for measuring derailment and incoherence.

This direction was developed further by Bar et al. [2019], who used `fastText` vectors [Bojanowski et al., 2016] to measure derailment in a study group that included 24 schizophrenia patients and 27 healthy controls, all native Hebrew speakers. Furthermore, they developed a new metric for measuring some aspects of incoherence, which compares the adjectives and adverbs that are used by patients to describe some nouns and verbs, respectively, with the ones used by the control group. As a final step, they used derailment and incoherence scores as features for training a classifier to separate the two study subgroups. In this work, we study a similar group of Hebrew-speaking male schizophrenia patients and healthy controls; therefore, we use some of the same linguistic features suggested in that prior work to measure their contribution when combined with acoustic features.

Chapter 3

Methodology

3.1 Participants and Data Collection

We interviewed 49 men, aged 18–60, divided into control and patient groups, all speaking Hebrew as their first language. The patient group includes 23 inpatients from the Be'er Ya'akov–Ness Ziona Mental Health Center in Israel who were admitted following a diagnosis of schizophrenia. Diagnoses were made by a hospital psychiatrist according to the DSM5 criteria (American Psychiatric Association, 2013) and a full psychiatric interview. Each participant was rewarded with approximately US$8. The control group includes 26 men, mainly recruited via an advertisement that we placed on social media. Exclusion criteria for all participants were as follows: (1) participants whose mother tongue is not Hebrew; (2) having a history of dependence on drugs or alcohol over the past year; (3) having a past or present neurological illness; and (4) using fewer than 500 words in total in their transcribed interview. Additionally, the control group had to score below the threshold for subclinical diagnosis of depression and post-traumatic stress disorder (PTSD). Most of the control participants scored below the threshold for anxiety. Most of the patients scored above the threshold for borderline or mild psychosis symptoms on a standard measure.[*] See Section 3.2 for more information about the measures we use in this study.

The demographic characteristics of the two groups are given in Table 3.1.

The patients were interviewed in a quiet room at the department where they are hospitalized by one of our professional team members, and the control participants were interviewed in a similar room outside the hospital. Each interview lasted approximately 60 minutes. The interviews were recorded and later manually transcribed by a native Hebrew speaking student from our lab. All participants were assured of anonymity, and told that they are free to end the interview at any time. Our research was approved by the Helsinki Ethical Review Board (IRB) of the Be'er Ya'akov-Ness Ziona Mental Health Center.

After signing a written consent, each participant was asked to describe 14 black and white images picked from the Thematic Appreciation Test (TAT) collection. We used the TAT images identified with the following

[*] Our patient group is composed of inpatients who are being treated through medications; therefore, higher scores were not expected.

Table 3.1: Demographic characteristics by group. *p<.05; **p<.005.

	Control	Patients	Statistics
Subjects (N)	25	23	
Age mean (SD)	33.15 (9.98)	25.46 (6.39)	t = 3.24**
Years of education mean (SD)	11.96 (0.20)	11.21 (1.12)	t = 3.41**
Place of residence (frequencies)			$\chi^2(3, 49) = 8.29$*
Southern Israel	1	7	
Central Israel	21	16	
Northern Israel	2	0	
Jerusalem	1	0	
Marital status (frequencies)			$\chi^2(1, 47) = 0.08$, p = .77
Single	4	3	
Married	21	20	
PANSS positive subscale		8.96 ± 3.85	
PANSS negative subscale		8.38 ± 3.91	
PANSS total subscale		17.34 ± 6.29	

serial numbers: 1, 2, 3BM, 4, 5, 6BM, 7GF, 8BM, 9BM, 12M, 13MF, 13B, 14, and 3GF. These include a mixture of men and women, children, and adults. The images were presented one by one. Each picture stands by itself, was presented alone, and bears no relation to the other pictures. Participants were asked to tell a brief story about each image based on four open questions:

(i) What led up to the event shown in the picture?
(ii) What is happening in the picture at this moment?
(iii) What are the characters thinking and feeling?
(iv) What is the outcome of the story?

The interviewer remained silent during the respondent's narration and offered no prompts or additional questions.

After describing the images, the participant was also asked to answer four open-ended questions, one by one:

(1) Please tell me as much as you can about your *bar mitzvah*.[†]
(2) What do you like to do, mostly?
(3) What are the things that annoy you the most?
(4) What would you like to do in the future?

As before, the interviewer remained silent during the respondent's narration and offered no prompts or questions.

Once all 18 components (14 image descriptions and 4 open questions) were answered, each participant was requested to fill in a demographic questionnaire as well as some additional questionnaires for assessing mental-health symptoms, which we describe in the following subsection.

[†] The Jewish confirmation ceremony for boys upon reaching the age of 13.

3.2 Symptom Assessment Measures

3.2.1 Control group

The control participants were assessed for symptoms of depression, PTSD, and anxiety.

Depression. Symptoms of depression were assessed using Beck's Depression Inventory-II (BDI-II) [Beck et al., 1996]. The BDI-II is a 21-item inventory rated on a 4-point Likert-type scale (0 = "not at all" to 3 = "extremely"), with summary scores ranging between 0 and 63. Beck et al. [1996] suggest a preliminary cutoff value of 14 as an indicator for mild depression, as well as a threshold of 19 as an indicator for moderate depression. BDI-II has been found to demonstrate high reliability [Gallagher et al., 1982]. We used a Hebrew version [Hasenson-Atzmon and Hermesh, 2016].

PTSD. Symptoms of PTSD were assessed using the PTSD checklist of the DSM-5 (PCL-5) [Weathers et al., 2013]. The questionnaire contains twenty items that can be divided into four subscales, corresponding to the clusters B-E in DSM-5: intrusion (five items), avoidance (two items), negative alterations in cognition and mood (seven items), and alterations in arousal and reactivity (six items). The items are rated on a 5-point Likert-type scale (0 = "not at all" to 4 = "extremely"). The total score ranges between 0 and 80, provided along with a preliminary cutoff score of 38 as an indicator for PTSD. PCL-5 has been found to demonstrate high reliability [Blevins et al., 2015]. We used a Hebrew translation of PCL-5 [Bensimon et al., 2013].

Anxiety. Symptoms of anxiety were assessed through the State Trait Anxiety Inventory (STAI) [Spielberger et al., 1970]. The STAI questionnaire consists of two sets of twenty self-reporting measures. The STAI measure of state anxiety (S-anxiety) assesses how respondents feel "right now, at this moment" (e.g., "I feel at ease"; "I feel upset"), and the STAI measure of trait anxiety (T-anxiety) targets how respondents "generally feel" (e.g., "I am a steady person"; "I lack self-confidence"). For each item, respondents are asked to rate themselves on a 4-point Likert scale, ranging from 1 = "not at all" to 4 = "very much so" for S-anxiety, and from 1 = "almost never" to 4 = "almost always" for T-anxiety. Total scores range from 20 to 80, with a preliminary cutoff score of 40 recommended as indicating clinically significant symptoms for the T-Anxiety scale [Knight et al., 1983]. STAI has been found to have high reliability [Barnes et al., 2002]. We used a Hebrew translation [Saka and Gati, 2007].

3.2.2 Patients

Psychosis symptoms were assessed by the 6-item Positive And Negative Syndrome Scale (PANSS-6) [Østergaard et al., 2016]. The original 30-item PANSS (PANSS-30) is the most widely used rating scale in schizophrenia, but it is relatively long for use in clinical settings. The items in PANSS-6 are rated on a 7-point scale (0 = "not at all" to 6 = "extremely"). The total score ranges from 0 to 36, with a score of

14 representing the threshold for mild schizophrenia, and a score between 10 and 14 defined as borderline disease or as remission. PANSS-30 has been found to demonstrate high reliability [Lin et al., 2018], while Østergaard et al. [2016] reported a high correlation between PANSS-6 and PANSS-30 (Spearman correlation coefficient = 0.86). We used the Hebrew version of PANSS-6 produced by Katz et al. [2012]. The range of positive and negative symptoms are presented in Table 1.

3.3 Data Analysis

We analyse the data using two modalities, audio and text. All the interviews were recorded with a voice recorder, which was placed on the table next to the participant. The responses of the participants for each of the 18 interview components were recorded separately, and stored as individual files in Waveform Audio File Format (WAV). Each response was manually transcribed. We extracted prosodic features from the audio signal, as well as linguistic features from the corresponding transcriptions.

3.3.1 Prosodic Acoustic Features

We processed each WAV file with PRAAT [Boersma, 2011], a computer software package for speech analysis, in order to extract pitch and intensity per 10ms frame. We distinguish between speech and non-speech frames by automatically annotating as speech those frames with a detected fundamental-frequency (F0) value below 250 Hz.

Each WAV file, corresponding to a response to a single image/question, is now represented by a sequence of speech frames, each represented by a pair of pitch and intensity values. We extract nine feature types from each response; to avoid overfitting, we filter out responses representing less than 10 seconds worth of speech. Therefore, we work with a dataset containing 449 responses given by controls and 409 responses given by patients. Following previous work on computational prosodic analysis [Kliper et al., 2015], we extracted the following set of features:

Mean Utterance Duration (MUD). Every segment of at least 500ms of continuous speech is defined as an *utterance*. *MUD* is the mean duration (in milliseconds) of all the utterances in a given response.

Mean Gap Duration (MGD). A *gap* is defined as a time interval containing no speech. *MGD* is the mean length (in milliseconds) of all gaps in a given response.

Mean Spoken Ratio (MSR). The sum of all utterances' duration divided by the total response duration.

Mean Spoken Ratio Samples (MSRS). The number of frames that are classified as speech, divided by the total number of frames in the response.

Table 3.2: Mean (SD) values of all prosodic features. *$p < .05$; ***$p < .001$.

Feature	**Control mean (SD)**	**Patient mean (SD)**	***t*-test**	p
MUD	0.798 (0.101)	0.629 (0.162)	4.143	< 0.001***
MGD	0.240 (0.066)	0.954 (0.990)	−3.602	< 0.001***
MSR	0.289 (0.118)	0.124 (0.087)	5.496	< 0.001***
MSRS	0.563 (0.100)	0.311 (0.138)	7.254	< 0.001***
MP	129.202 (22.113)	125.422 (21.274)	0.602	0.550
PR	1.148 (0.153)	0.906 (0.196)	4.809	< 0.001***
PS	21.579 (5.517)	19.411 (7.370)	1.160	0.252
MWC	0.581 (0.096)	0.483 (0.147)	2.750	0.009*
J	0.008 (0.002)	0.007 (0.002)	2.066	0.044*

Mean Pitch (MP). The mean pitch values (in Hz) of all frames recognized as speech in a given response.

Pitch Range (PR). The maximum value of pitch of all frames recognized as speech, minus their minimum value, and divided by MP for normalization. It is measured in Hz.

Pitch Std (PS). The standard deviation of pitch values (in Hz) of all frames recognized as speech in a given file.

Mean Waveform Correlation (MWC). The Pearson correlation between a sequence of pitch values of speech frames and a sequence of pitch values of their consecutive frames.

Jitter (J). The local deviation from stationarity of the pitch. Formally, let R be the number of speech frames, and let $p(v)$ be the pitch value of the vth frame. We define J as follows:

$$\mathrm{J} := \frac{1}{R-K} \sum_{v=\frac{K-1}{2}}^{R-\frac{K-1}{2}-1} \frac{p(v) - \frac{1}{K}\sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} p(v+k)}{\sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} p(v+k)}$$

K is a locality parameter; it was set to 5 in all our experiments.

We did not extract features that are based on intensity, since we noticed some differences in the background noise between the recordings of the control participants and the patients, probably due to differences in room settings and recording equipment.

We verified that all our features are distributed normally, as expected, and performed t-tests to measure the difference in feature expression between patients and controls. The results are summarized in Table 3.2. As can be seen, most features are distributed significantly differently for patients and controls. However, the mean and standard deviation of the pitch values seem to be similar in the two groups.

3.3.2 Linguistic Features

We extract the same linguistic features that have been used by Bar et al. [2019] on a similar dataset. Essentially, they designed two types of features for capturing specific symptoms of thought disorder.

Derailment. The first type is designed to capture derailment, which is a symptom of thought disorder when the speaker digresses from the main topic. Technically speaking, we represent words using static embeddings provided by `fastText` [Grave et al., 2018] for Hebrew. For each response, we retrieve the `fastText` vector v_i for every word R_i, $i = 0..n$, in the response. Then, for each word, we calculate a score defined as the average pairwise cosine similarity between this word and the k following words, with k a variable parameter. The score of a response is the average of all the individual cosine-similarity scores. To filter out functional words that do not contribute to the topical mutation assessment, we follow [Bar et al., 2019] by pre-processing each response with a Hebrew part-of-speech tagger [Adler, 2007] and keep only content words, which we take to be nouns, verbs, adjectives, and adverbs.

We calculate derailments for $k = 1..6$, thereby extracting six derailment features per response.

Incoherence. One of the most informative features reported in [Bar et al., 2019] was designed to capture some aspects of discourse related to incoherence. Specifically, this feature examines the way patients use adjectives to describe specific nouns. The goal is to measure the difference between adjectives used by patients and the ones used by controls when describing the same nouns. Technically speaking, we process each response with `YAP` [More and Tsarfaty, 2016], a dependency parser for Modern Hebrew, to find all noun-adjective pairs (indicated by the *amod* relation). To measure the difference between adjectives that are used by patients and controls, we compare them to the adjectives that are commonly used to describe the same nouns and verbs. To do that, following the above-mentioned work, we use an external corpus of health-related documents and forums, all written in Hebrew, containing nearly 680K words.[‡] We process each document in exactly the same way to find all noun-adjective pairs. Given a list of noun-adjective pairs from one response, we calculate the similarity score between every adjective that describes a specific noun and the set of adjectives describing exactly the same noun across the entire external corpus. Hebrew enjoys a rich morphology; therefore, we work on the lemma (base-form) level. The lemmata are provided by `YAP`. We take the `fastText` vectors of the adjectives that were extracted from the external corpus and average them, element wise, into a single vector by assigning weights to each individual vector. The weights are the inverse-document-frequency (idf) score of each adjective, to account more heavily for adjectives that describe the noun more uniquely. Then, we take the cosine similarity between each adjective from the response and the aggregated vector of the adjectives from the external corpus. For each response, we take the average of the individual adjective cosine-similarity scores as the overall response incoherence score.

As before, we verified that all our features are distributed normally and performed t-tests to measure the difference in feature expression between patients and controls. The results are summarized in Table 3.3.

[‡] We use exactly the same sources used in [Bar et al., 2019].

Table 3.3: Mean (SD) values of the linguistic features.

Feature	**Control mean (SD)**	**Patient mean (SD)**	***t*-test**	*p*
Derailment 1	0.247 (0.011)	0.239 (0.017)	1.797	0.080
Derailment 2	0.237 (0.015)	0.236 (0.013)	0.102	0.918
Derailment 3	0.233 (0.015)	0.231 (0.017)	0.297	0.768
Derailment 4	0.229 (0.015)	0.226 (0.021)	0.522	0.605
Derailment 5	0.227 (0.016)	0.226 (0.016)	0.331	0.742
Derailment 6	0.225 (0.016)	0.225 (0.016)	0.006	0.995
Incoherence	0.520 (0.062)	0.502 (0.070)	0.931	0.357

In contrast with the outcomes in [Bar et al., 2019], we see no evidence for different distributions of each individual linguistic feature between the two groups.

3.4 Classification

We train a two-way machine-learning classifier to distinguish between responses that were generated by patients and controls. Each response is used as a classification instance, assigned either a patient or control label depending on the group to which the subject who generated the response belongs. Overall we have 449 responses generated by controls and 409 responses by patients. We ran three sets of experiments: (1) using only the acoustic features (Acoustic); (2) using only the linguistic features (Linguistic); and, (3) using both feature sets (Combined). Consequently, each response is represented by a nine-dimensional vector in the first set of experiments, a seven-dimensional vector in the second set, and a 16-dimensional vector in the third set of experiments.

For classification, we used three traditional machine-learning algorithms: XGBoost [Chen and Guestrin, 2016], Random Forest [Liaw et al., 2002], and Linear SVM [Cortes and Vapnik, 1995].

Chapter 4

Results

We measured the classification results using accuracy and the F1 score of the patient label. For each classifier, we ran five evaluations, each time taking a five-fold cross-validation approach. Every evaluation had a different random seed, which was kept similar across all classifiers. The five results were calculated as the average over the five evaluation runs. The results, divided into the three feature sets, are presented in Table 4.1.

Overall XGBoost delivers the best classification results. It reaches the highest accuracy and F1 score when using the set of acoustic features. When using only the linguistic features, all the classifiers perform more poorly. Furthermore, combining the linguistic features with the acoustic ones did not result in performance improvement, suggesting that the contribution of the linguistic features to the classification performance on our dataset is limited and redundant when pitch-based acoustic features are used for detecting symptoms of schizophrenia. The lesser success with linguistic features may in part be due to the inherent difficulty of accurately measuring semantic features like derailment and incoherence computationally.

Our best accuracy for the two-way classification task is around 90%, which is higher than the best

Table 4.1: Classification results.

Feature Set	**Classifier**	**Accuracy (SD)**	**Precision (SD)**	**Recall (SD)**	**F1 (SD)**
Acoustic	Random Forest	86.8 (4.2)	82.6 (5.3)	85.0 (5.4)	82.2 (4.7)
Acoustic	XGBoost	**91.4 (0.7)**	**97.0 (0.4)**	86.0 (2.7)	88.9 (1.4)
Acoustic	Linear SVM	88.1 (3.0)	91.1 (2.9)	83.0 (4.5)	84.8 (3.5)
Linguistic	Random Forest	65.0 (4.7)	66.5 (5.6)	45.9 (4.2)	51.4 (4.6)
Linguistic	XGBoost	63.1 (4.8)	64.3 (5.7)	60.6 (4.8)	59.5 (5.4)
Linguistic	Linear SVM	73.4 (2.7)	74.0 (5.0)	65.3 (4.9)	66.3 (3.6)
Combined	Random Forest	90.7 (2.0)	94.5 (1.6)	**89.6 (2.8)**	**90.0 (1.8)**
Combined	XGBoost	88.5 (1.0)	92.9 (3.5)	82.6 (3.4)	84.9 (2.5)
Combined	Linear SVM	88.4 (2.0)	87.9 (4.6)	82.0 (3.7)	83.4 (3.8)

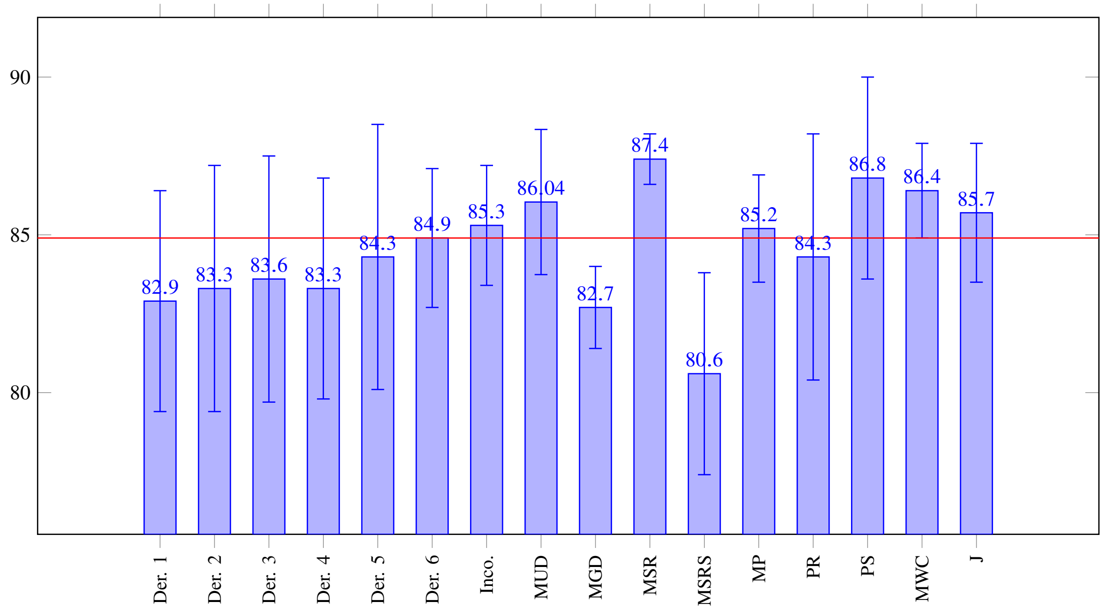

Figure 4.1: Ablation study: F1 (y axis) scores of the Combined XGBoost classifier by removing one feature from the data at a time, as indicated by the x axis. Der. = Derailment; Inco. = Incoherence.

accuracy of about 76% reported by Kliper et al. [2015] using a similar set of acoustic features for the same two-way classification task with an English-speaking population.

To measure the correlation between all the individual features, we calculate Pearson ρ values, and show the as a heat map in Figure 4.2. Unsurprisingly, we see a strong correlation between all the linguistic derailment features, which make them somewhat redundant for classification. Among the acoustic features, we see a stronger correlation between the standard deviation of the pitch and the mean waveform correlation. Generally speaking, both represent the dynamics of the pitch in speech frames. Similarly, and unsurprisingly, mean spoken ratio is strongly correlated with mean spoke ratio samples; both represent the ratio between the time in which actual speaking is taking place, and the overall time of the response. Naturally, mean gap duration has a negative correlation with all the features that measure speaking duration. However, we do not see any significant correlation between the acoustic features and the linguistic ones. Nonetheless, as seen in Table 4.1, the linguistic features did not contribute new information for classification not already covered by the acoustic features.

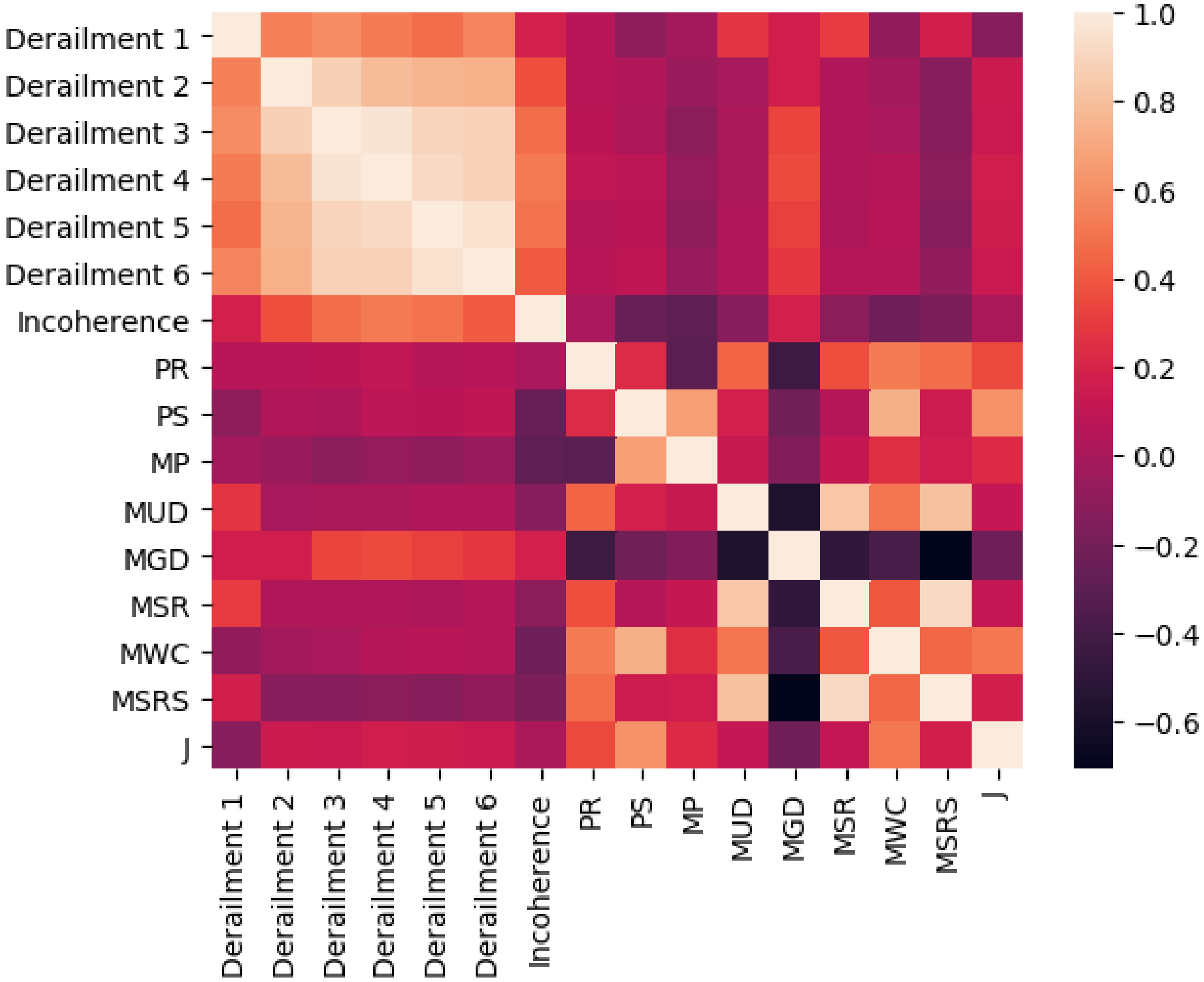

Figure 4.2: Pearson ρ between all individual features, presented as a heat map.

Chapter 5

Conclusion

We have extracted features from two modalities of Hebrew speech produced by schizophrenia patients during interviews and compared it with those of controls. Specifically, we extracted acoustic, prosodic features from the audio signal, as well as linguistic features of transcriptions of the interview that measure derailment and incoherence. Our main goal was to measure the contribution of each modality to classification performance, when used in combination. Generally speaking, we find that a traditional classification algorithm can nicely separate between the two groups, schizophrenia patients and healthy controls, with best accuracy of about 90%, which is better than the results that have been previously reported. The results also show that the linguistic features do not add much to classification performance when they are combined with the acoustic features that measure aspects of prosody.

Bibliography

Meni Adler. *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. PhD thesis, Ben-Gurion University of the Negev, Beer-Sheva, Israel, 2007.

Carla Agurto, Mary Pietrowicz, Raquel Norel, Elif K. Eyigoz, Emma Stanislawski, Guillermo Cecchi, and Cheryl Corcoran. Analyzing acoustic and prosodic fluctuations in free speech to predict psychosis onset in high-risk youths. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5575–5579. IEEE, 2020.

Mark S. Aloia, Monica L. Gourovitch, David Missar, David Pickar, Daniel R. Weinberger, and Terry E. Goldberg. Cognitive substrates of thought disorder, II: Specifying a candidate cognitive mechanism. *American Journal of Psychiatry*, 155(12):1677–1684, 1998.

Murray Alpert, Stanley D. Rosenberg, Enrique R. Pouget, and Richard J. Shaw. Prosody and lexical accuracy in flat affect schizophrenia. *Psychiatry Research*, 97(2):107–118, 2000. ISSN 0165-1781. doi: 10.1016/S0165-1781(00)00231-6.

DS American Psychiatric Association, American Psychiatric Association, et al. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, volume 5. American Psychiatric Association, Washington, DC, 2013.

Nancy C. Andreasen. Thought, language, and communication disorders: I. clinical assessment, definition of terms, and evaluation of their reliability. *Archives of General Psychiatry*, 36(12):1315–1321, 1979.

Kfir Bar, Vered Zilberstein, Ido Ziv, Heli Baram, Nachum Dershowitz, Samuel Itzikowitz, and Eiran Vadim Harel. Semantic characteristics of schizophrenic speech. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 84–93, Minneapolis, MN, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3010. URL `https://aclanthology.org/W19-3010`.

Laura L. B. Barnes, Diane Harp, and Woo Sik Jung. Reliability generalization of scores on the Spielberger state-trait anxiety inventory. *Educational and Psychological Measurement*, 62(4):603–618, 2002.

Aaron T. Beck, Robert A. Steer, Roberta Ball, and William F. Ranieri. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment*, 67(3):588–597, 1996.

Gillinder Bedi, Facundo Carrillo, Guillermo A. Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B. Mota, Sidarta Ribeiro, Daniel C. Javitt, Mauro Copelli, and Cheryl M. Corcoran. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1:15030, 2015.

Moshe Bensimon, Stephen Zvi Levine, Gadi Zerach, Einat Stein, Vlad Svetlicky, and Zahava Solomon. Elaboration on posttraumatic stress disorder diagnostic criteria: A factor analytic study of PTSD exposure to war or terror. *Israel Journal of Psychiatry*, 50(2):84–90, 2013.

Christy A. Blevins, Frank W. Weathers, Margaret T. Davis, Tracy K. Witte, and Jessica L. Domino. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, 28(6):489–498, 2015.

Paul Boersma. Praat: doing phonetics by computer, 2011. URL `http://www.praat.org`. Computer program.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. URL `https://arxiv.org/pdf/1607.04606.pdf`.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Colin Cherry. page 294. J. & A. Churchill, Ltd., London, 1964.

Alex S. Cohen, Yunjung Kim, and Gina M. Najolia. Psychiatric symptom versus neurocognitive correlates of diminished expressivity in schizophrenia and mood disorders. *Schizophrenia Research*, 146(1–3): 249–253, 2013.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Michael A. Covington, Congzhou He, Cati Brown, Lorina Naçi, Jonathan T. McClain, Bess Sirmon Fjordbak, James Semple, and John Brown. Schizophrenia and the structure of language: The linguist's view. *Schizophrenia Research*, 77(1):85–98, 2005. ISSN 0920-9964. doi: 10.1016/j.schres.2005.01.016.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391–407, 1990.

Chandlee C. Dickey, Mai-Anh T. Vu, Martina M. Voglmaier, Margaret A. Niznikiewicz, Robert W. McCarley, and Lawrence P. Panych. Prosodic abnormalities in schizotypal personality disorder. *Schizophrenia Research*, 142(1–3):20–30, 2012.

Brita Elvevåg, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93 (1–3):304–316, 2007.

Dolores Gallagher, Gloria Nies, and Larry W. Thompson. Reliability of the beck depression inventory with older adults. *Journal of Consulting and Clinical Psychology*, 50(1):152–153, February 1982.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Kelly Hasenson-Atzmon and Haggai Hermesh. Cultural impact on SAD: Social anxiety disorder among Ethiopian and former Soviet Union immigrants to Israel, in comparison to native-born Israelis. *Israel Journal of Psychiatry*, 53(3):48–54, 2016.

Fei He, Ling He, Jing Zhang, Yuan yuan Li, and Xi Xiong. Automatic detection of affective flattening in schizophrenia: Acoustic correlates to sound waves and auditory perception. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

Marjolijn Hoekert, René S. Kahn, Marieke Pijnenborg, and André Aleman. Impaired recognition and expression of emotional prosody in schizophrenia: Review and meta-analysis. *Schizophrenia Research*, 96(1):135–145, 2007. ISSN 0920-9964. doi: 10.1016/j.schres.2007.07.023.

Yan-Jia Huang, Yi-Ting Lin, Chen-Chung Liu, Lue-En Lee, Shu-Hui Hung, Jun-Kai Lo, and Li-Chen Fu. Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:947–956, 2022.

Dan Iter, Jong Yoon, and Dan Jurafsky. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146, 2018.

Gregory Katz, Leon Grunhaus, Shukrallah Deeb, Emi Shufman, Rachel Bar-Hamburger, and Rimona Durst. A comparative study of arab and jewish patients admitted for psychiatric hospitalization in jerusalem: The demographic, psychopathologic aspects, and the drug abuse comorbidity. *Comprehensive Psychiatry*, 53 (6):850–853, 2012.

Roi Kliper, Yonatan Vaizman, Daphna Weinshall, and Shirley Portuguese. Evidence for depression and schizophrenia in speech prosody. In *Proceedings of the Third ISCA Workshop on Experimental Linguistics*, pages 35–38, August 2010. URL https://www.isca-speech.org/archive_v0/exling_2010/papers/el10_085.pdf.

Roi Kliper, Shirley Portuguese, and Daphna Weinshall. Prosodic analysis of speech and the underlying mental state. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 52–62. Springer, 2015.

Robert G. Knight, Hendrika J. Waal-Manning, and George F. Spears. Some norms and reliability data for the state-trait anxiety inventory and the Zung self-rating depression scale. *British Journal of Clinical Psychology*, 22(4):245–249, 1983.

Andy Liaw, Matthew Wiener, et al. Classification and regression by randomForest. *R News*, 2(3):18–22, December 2002.

C.-H. Lin, H.-S. Lin, S.-C. Lin, C.-C. Kuo, F.-C. Wang, and Y.-H. Huang. Early improvement in PANSS-30, PANSS-8, and PANSS-6 scores predicts ultimate response and remission during acute treatment of schizophrenia. *Acta Psychiatrica Scandinavica*, 137(2):98–108, 2018.

Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116, 2020.

Francisco Martínez-Sánchez, José Antonio Muela-Martínez, Pedro Cortés-Soto, Juan José García Meilán, Juan Antonio Vera Ferrándiz, Amaro Egea Caparrós, and Isabel María Pujante Valverde. Can the acoustic analysis of expressive prosody discriminate schizophrenia? *The Spanish Journal of Psychology*, 18, November 2015. Article E86.

Amir More and Reut Tsarfaty. Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING 2016*, December 2016.

Noa Saka and Itamar Gati. Emotional and personality-related aspects of persistent career decision-making difficulties. *Journal of Vocational Behavior*, 71(3):340–358, 2007.

Charles Donald Spielberger, Richard L. Gorsuch, and Robert E. Lushene. *STAI Manual for the State-Trait Anxiety Inventory ("self-evaluation questionnaire")*. Consulting Psychologists Press, Palo Alto, 1970.

T. H. Spoerri. Speaking voice of the schizophrenic patient. *Archives of General Psychiatry*, 14(6):581–585, 1966.

Frank W. Weathers, Brett T. Litz, Terence M. Keane, Patrick A. Palmieri, Brian P. Marx, and Paula P. Schnurr. The PTSD checklist for DSM-5 (PCL-5), 2013. Scale available from the National Center for PTSD at www.ptsd.va.gov.

Soren Dinesen Østergaard, Ole Michael Lemming, Ole Mors, Christoph U. Correll, and Per Bech. PANSS-6: A brief rating scale for the measurement of severity in schizophrenia. *Acta Psychiatrica Scandinavica*, 133(6):436–444, 2016.

הפקולטה למדעים
מדויקים ע"ש ריימונד
ובברלי סאקלר
אוניברסיטת תל אביב

התרומה של פרוזודיה על זיהוי ממוחשב של סכיזופרניה

חיבור זה

הוגש כחלק מהדרישות לקבלת תואר

מוסמך אוניברסיטה

על ידי

תומר בן משה

העבודה הוכנה בהנחיית

ד"ר כפיר בר

פרופ' נחום דרשוביץ

אוניברסיטת תל אביב

בית הספר למדעי המחשב

חשון ה'תשפ"ג

תקציר:

אנחנו מראים איך מאפיינים פרוזודיים, כגון גובה טון ואורך הברה בדיבור, יכולים לעזור למחשב לזהות סימפטומים של סכיזופרניה מתגובה של אדם בודד.

אנחנו משווים את התרומה של מודלי דיבור וטקסט על הזיהוי האם לאדם נתון יש סכיזופרניה.

התוצאות שלנו מראות בצורה מובהקת שמאפיינים פרוזודיים שובים יותר מידע מאשר מאפיינים הקשורים לתוכן של הדיבור(מאפיינים שפתיים).

אנחנו מוצאים שכאשר משווים למאפיינים פרוזודיים, המאפיינים השפתיים תורמים רק במעט לזיהוי סכיזופרניה.