The Contribution of Prosody to Machine Classification of Schizophrenia

Tomer Ben Moshe Ido Ziv Nachum Dershowitz
Kfir Bar

September 6, 2023

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 previously-employed text modalities to the algorithmic 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.

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 are 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, shifting from one topic to another during a conversation; and (5) tangentiality, furnishing an irrelevant response, never reaching the answer to the posed question. Andreasen (1979) provides

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 several prior 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 human language that can be processed and analyzed algorithmically 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 prominent 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 occasionally 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 required for capturing linguistic and semantic characteristics of conversation.

We study the salience of prosodic (speech-based) and linguistic (textual) features for the classification task of automatically detecting whether a given utterance was generated by someone who has been diagnosed with schizophrenia or by a control subject. To do that, we measure the contribution of each set of features once when used individually for classification, and again when both modalities 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 consider are based on pitch, which we extract using an audio processor. The linguistic features are extracted from the transcriptions 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 the task of detecting schizophrenia—for example by Kliper, Vaizman, et al. (2010) and Kliper, Portuguese, et al. (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. Our classification results clearly show that the acoustic prosodic features capture more information than do the linguistic ones. When combined with those prosodic features, linguistic features improve classification very slightly.

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 assert 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 seventeen 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 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, Portuguese, 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 versus 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 similitude 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 perfect 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 another work

(Ziv et al., 2022) on the same study group, the authors used part-of-speech tags, lemma-to-token ratio, and some other morphological features, to perform a two-way classification for patients and controls. They report on almost 90% accuracy.

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 respective contributions when combined with acoustic features.

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 25 men, mainly recruited via an advertisement that we placed on social media. The demographic characteristics of the two groups are given in Table 1. 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. (Our patient group is composed of inpatients who are being treated with medications; therefore, higher scores were not expected.) See Section 3.2 for more details about the assessment measures used in this study.

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

Table 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)			$\chi^2(1, 47) = 0.08$,
Single	4	3	p = .77
Married	21	20	
PANSS positive subscale		8.96 ± 3.85	
PANSS negative subscale		8.38 ± 3.91	
PANSS total subscale		17.34 ± 6.29	

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.

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 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*.[1]
(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.

NB. This research was approved by the Helsinki Ethical Review Board (IRB) of the Be'er Ya'akov–Ness Ziona Mental Health Center.

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 et al., 2016).

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

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 the last three rows of 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, Portuguese, 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 ms) of all the utterances in a given response.

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

Mean Spoken Ratio (MSR). The sum of the durations of all utterances 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.

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

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

Standard Deviation of Pitch in a Single Response (PS). The standard deviation of pitch (in Hz) of all frames recognized as speech in a given response.

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

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

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

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

Table 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*

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 2. As can be seen, most features are distributed significantly differently for patients and controls. However, the mean and standard deviation of the pitch of a given response (MP and PS) seem to be similar in the two groups, suggesting that the analysis over the pitch, made by the other features, is important for distinguishing between the two groups. It is worth noting that among the patient group, MGD exhibits relatively high levels of variability, as indicated by the relatively large standard deviation.

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.[2] 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

[2]We use the same sources as in (Bar et al., 2019).

Table 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

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. 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]

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.

[3]The datasets, patients and controls, differ for the two experiments. And, in the previous work, the controls were told to talk for at least two minutes, which potentially impacted the outcome.

Table 4: 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)

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

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.

Overall, the XGBoost algorithm achieves the best classification accuracy when utilizing solely the acoustic features. On the other hand, Random Forest achieves the best F1 score using all features. When using only the linguistic features, all the classifiers perform poorly. Furthermore, combining the linguistic features with the acoustic ones did not usually result in significant performance improvement, suggesting that the contribution of our linguistic features to the classification performance on the dataset is limited and redundant when pitch-based acoustic features are used for detecting symptoms of schizophrenia. The lesser success

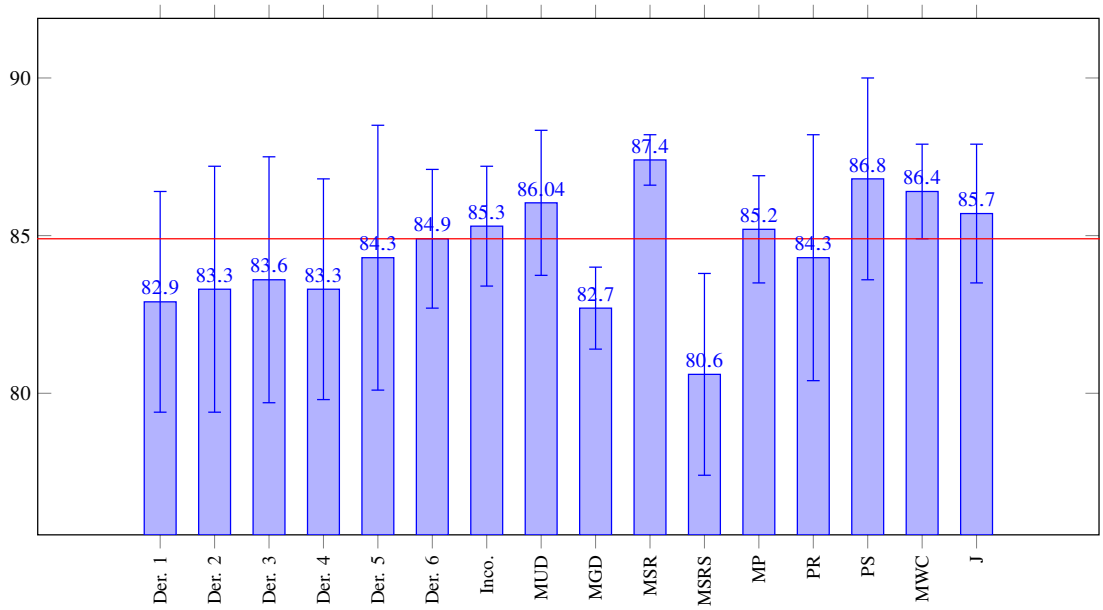

Figure 1: Ablation study: F1 (y axis) scores of the Combined XGBoost classifier by removing one feature from the data at a time, as indicated by the x axis. The red line at 84.9 indicates the F1 value for XGBoost with all features included. The F1 scores the average of five executions, each using a different seed. Der. = Derailment; Inco. = Incoherence.

with linguistic features may be due in part 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 accuracy of about 76% reported by **Rkliper2015prosodic** using a similar set of acoustic features for the same two-way classification task with an English-speaking population.

Looking at the demographic characteristics of the participants in Table 1, we notice that the patients and controls significantly differ in age and years of education. Therefore, we performed a complementary analysis to support the current findings in which seven sets of multiple regressions have been carried out as reported in Table 5. These represent the seven prosodic features which demonstrated a significant (at least $p < .05$) different distribution among patients and controls. As shown in Table 3, none of the linguistic features has been shown

to be different among the two groups.

As can seen from Table 5, years of education consistently did not associate with any of the prosodic features. The age characteristic was associated significantly ($p = .047$) only once with MGD. However, the group (patients/control) was the only predictor that was associated consistently, substantially, and significantly with all the prosodic features.

We performed an ablation study to measure the effect of each feature individually. The results are summarized in Figure 1. As can be seen, MSRS and MGD are the most effective features; both are related to the pace of speech. It is noteworthy that removing certain features, primarily acoustic ones, slightly improves the performance of the classifier. The most significant one is MSR, which is related to the total duration of speech with respect to the overall response duration. Our hypothesis is that this is mainly a result of overlap in our feature descriptions, as MSR, MGD, and MSRS are slight variations of the same idea to some extent.

To measure the correlation between all the individual features, we calculate Pearson ρ for all feature pairs and summarize them in a heat map, as shown in Figure 2. Unsurprisingly, we see a strong correlation between all the linguistic derailment features, which makes them somewhat redundant for classification. Among the acoustic features, we see a stronger correlation between the standard deviation of the pitch (SP) and the mean waveform correlation (MWC). Generally speaking, both represent the dynamics of the pitch in speech frames. Similarly, and unsurprisingly, the mean spoken ratio (MSR) is strongly correlated with mean spoken ratio samples (MSRS); both represent the ratio between the time in which actual speaking is taking place and the overall time of the response. Naturally, gap duration (MGD) has a negative correlation with all the features that measure speaking duration. However, we do not find any significant correlation between the acoustic features and the linguistic ones. And, as seen in Table 4, the linguistic features did not contribute added information for classification not already covered by the acoustic prosodic features.

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.

Table 5: Seven regression analyses for the most impacting prosodic features. PEF = percentage of explained variance; Edu. = Years of education; Grp. = Group (patients/control). For more information, see the text. $*p < .05$, $**p < .01$, $***p < .001$

	MUD		**MGD**		**MSRS**		**MSR**		**PR**		**MWC**		**J**	
	β	t	β	t	β	t	β	t	β	t	β	t	β	t
Age	-.04	.34	-.29	2.04*	.05	.49	-.08	.65	.24	1.88	.16	1.04	.18	1.1
Edu.	.06	.46	-.05	.38	.07	.62	.13	1.02	-.17	1.30	-.10	.68	-.06	.42
Grp.	.47	3.04**	.57	3.76***	.72	5.87***	.53	3.85***	.76	5.4***	.49	2.98**	.4	2.35*
PEV	$R^2 = .28$ $F(3,45) =$ 5.63**		$R^2 = .29$ $F(3,45) =$ 5.95**		$R^2 = .54$ $F(3,45) =$ 17.20***		$R^2 = .41$ $F(3,45) =$ 10.4***		$R^2 = .41$ $F(3,45) =$ 10.19***		$R^2 = .17$ $F(3,45) =$ = 3.05*		$R^2 = .11$ $F(3,45)$ = 1.93	

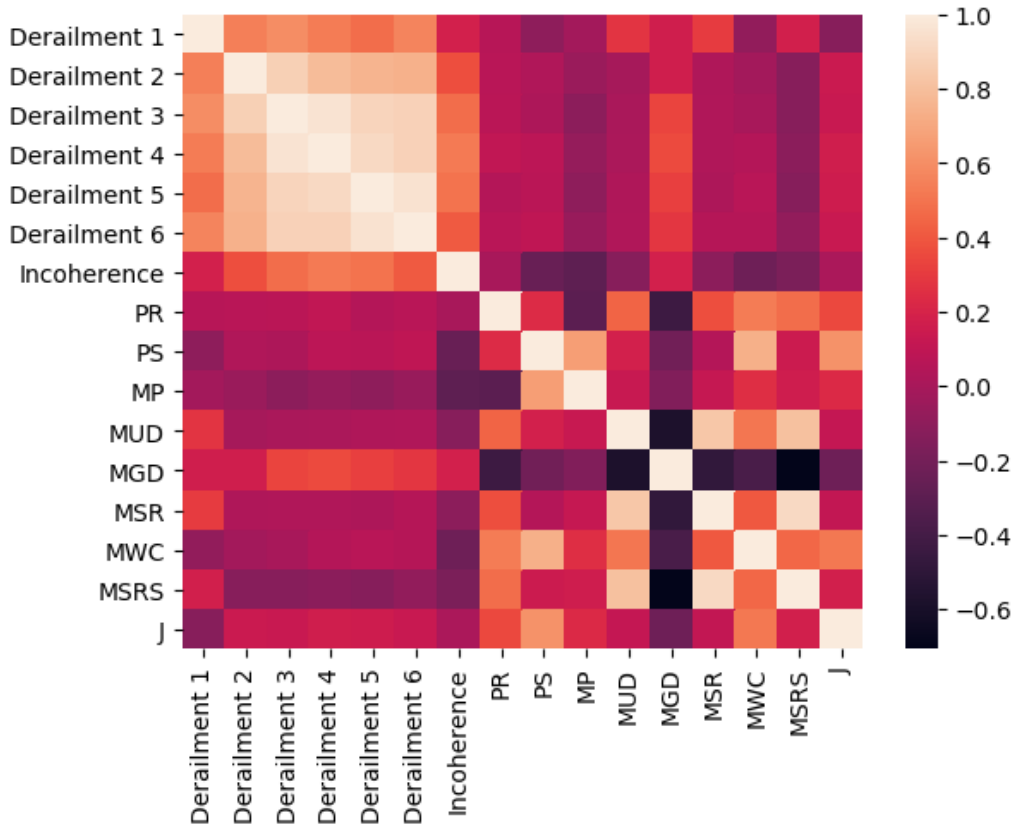

Figure 2: Pearson ρ between all individual features, shown as a heat map.

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.

References

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

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

Aloia, Mark S., Monica L. Gourovitch, David Missar, David Pickar, Daniel R. Weinberger, and Terry E. Goldberg (1998). "Cognitive substrates of thought disorder, II: Specifying a candidate cognitive mechanism". In: *American Journal of Psychiatry* 155.12, pp. 1677–1684.

Alpert, Murray, Stanley D. Rosenberg, Enrique R. Pouget, and Richard J. Shaw (2000). "Prosody and lexical accuracy in flat affect schizophrenia". In: *Psychiatry Research* 97.2, pp. 107–118. ISSN: 0165-1781. DOI: 10.1016/S0165-1781(00)00231-6.

American Psychiatric Association DSM-5 Task Force (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Vol. 5. Washington, DC: American Psychiatric Publishing.

Andreasen, Nancy C. (1979). "Thought, language, and communication disorders: II. Diagnostic significance". In: *Archives of General Psychiatry* 36.12, pp. 1325–1330.

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

Barnes, Laura L. B., Diane Harp, and Woo Sik Jung (2002). "Reliability generalization of scores on the Spielberger state-trait anxiety inventory". In: *Educational and Psychological Measurement* 62.4, pp. 603–618.

Beck, Aaron T., Robert A. Steer, Roberta Ball, and William F. Ranieri (1996). "Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients". In: *Journal of Personality Assessment* 67.3, pp. 588–597. DOI: 10.1207/s15327752jpa6703_13.

Bedi, Gillinder, Facundo Carrillo, Guillermo Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália Mota, Sidarta Ribeiro, Daniel Javitt, Mauro Copelli,

and Cheryl Corcoran (Sept. 2015). "Automated analysis of free speech predicts psychosis onset in high-risk youths". In: *npj Schizophrenia* 1. Article 15030. DOI: 10.1038/npjschz.2015.30.

Bensimon, Moshe, Stephen Zvi Levine, Gadi Zerach, Einat Stein, Vlad Svetlicky, and Zahava Solomon (2013). "Elaboration on posttraumatic stress disorder diagnostic criteria: A factor analytic study of PTSD exposure to war or terror". In: *Israel Journal of Psychiatry* 50.2, pp. 84–90.

Blevins, Christy A., Frank W. Weathers, Margaret T. Davis, Tracy K. Witte, and Jessica L. Domino (2015). "The posttraumatic stress disorder checklist for *DSM-5* (PCL-5): Development and initial psychometric evaluation". In: *Journal of Traumatic Stress* 28.6, pp. 489–498.

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

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

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

Cherry, Colin (1964). In: *Disorders of Language: Ciba Foundation Symposium*. Ed. by A. V. S. de Reuck and Maeve O'Connor. London: J. & A. Churchill, Ltd., p. 294.

Cohen, Alex S., Yunjung Kim, and Gina M. Najolia (2013). "Psychiatric symptom versus neurocognitive correlates of diminished expressivity in schizophrenia and mood disorders". In: *Schizophrenia Research* 146.1–3, pp. 249–253.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine Learning* 20.3, pp. 273–297.

Covington, Michael A., Congzhou He, Cati Brown, Lorina Naçi, Jonathan T. McClain, Bess Sirmon Fjordbak, James Semple, and John Brown (2005). "Schizophrenia and the structure of language: The linguist's view". In: *Schizophrenia Research* 77.1, pp. 85–98. ISSN: 0920-9964. DOI: 10.1016/j.schres.2005.01.016.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990). "Indexing by latent semantic analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Dickey, Chandlee C., Mai-Anh T. Vu, Martina M. Voglmaier, Margaret A. Niznikiewicz, Robert W. McCarley, and Lawrence P. Panych (2012). "Prosodic

abnormalities in schizotypal personality disorder". In: *Schizophrenia Research* 142.1–3, pp. 20–30.

Elvevåg, Brita, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg (2007). "Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia". In: *Schizophrenia Research* 93.1–3, pp. 304–316.

Gallagher, Dolores, Gloria Nies, and Larry W. Thompson (1982). "Reliability of the Beck Depression Inventory with older adults". In: *Journal of Consulting and Clinical Psychology* 50.1, pp. 152–153.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov (2018). "Learning word vectors for 157 languages". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 3483–3487.

Hasenson-Atzmon, Kelly, Sofi Marom, Tamar Sofer, Lilac Lev-Ari, Rafael Youngmann, Haggai Hermesh, Jonathan Kushnir, and Haggai Hermesh (2016). "Cultural impact on SAD: Social anxiety disorder among Ethiopian and Former Soviet Union immigrants to Israel, in comparison to native-born Israelis". In: *Israel Journal of Psychiatry* 53.3, pp. 48–54.

He, Fei, Ling He, Jing Zhang, Yuan Yuan Li, and Xi Xiong (2021). "Automatic detection of affective flattening in schizophrenia: Acoustic correlates to sound waves and auditory perception". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 3321–3334. DOI: 10.1109/TASLP.2021.3120591.

Hoekert, Marjolijn, René S. Kahn, Marieke Pijnenborg, and André Aleman (2007). "Impaired recognition and expression of emotional prosody in schizophrenia: Review and meta-analysis". In: *Schizophrenia Research* 96.1, pp. 135–145. ISSN: 0920-9964. DOI: 10.1016/j.schres.2007.07.023.

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

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

Katz, Gregory, Leon Grunhaus, Shukrallah Deeb, Emi Shufman, Rachel Bar-Hamburger, and Rimona Durst (2012). "A comparative study of Arab and Jewish patients admitted for psychiatric hospitalization in Jerusalem: The de-

mographic, psychopathologic aspects, and the drug abuse comorbidity". In: *Comprehensive Psychiatry* 53.6, pp. 850–853.

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

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

Knight, Robert G., Hendrika J. Waal-Manning, and George F. Spears (1983). "Some norms and reliability data for the State-Trait Anxiety Inventory and the Zung Self-Rating Depression scale". In: *British Journal of Clinical Psychology* 22.4, pp. 245–249.

Liaw, Andy, Matthew Wiener, et al. (Dec. 2002). "Classification and regression by randomForest". In: *R News* 2.3, pp. 18–22.

Lin, Ching-Hua, Huey-Shyan Lin, Shih-Chi Lin, Chao-Chan Kuo, Fu-Chiang Wang, and Yu-Hui Huang (2018). "Early improvement in PANSS-30, PANSS-8, and PANSS-6 scores predicts ultimate response and remission during acute treatment of schizophrenia". In: *Acta Psychiatrica Scandinavica* 137.2, pp. 98–108.

Low, Daniel M., Kate H. Bentley, and Satrajit S. Ghosh (2020). "Automated assessment of psychiatric disorders using speech: A systematic review". In: *Laryngoscope Investigative Otolaryngology* 5.1, pp. 96–116.

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

More, Amir and Reut Tsarfaty (Dec. 2016). "Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies". In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. Osaka, Japan, pp. 337–348. URL: https://aclanthology.org/C16-1033.pdf.

Østergaard, Soren Dinesen, Ole Michael Lemming, Ole Mors, Christoph U. Correll, and Per Bech (2016). "PANSS-6: A brief rating scale for the measurement of severity in schizophrenia". In: *Acta Psychiatrica Scandinavica* 133.6, pp. 436–444.

Saka, Noa and Itamar Gati (2007). "Emotional and personality-related aspects of persistent career decision-making difficulties". In: *Journal of Vocational Behavior* 71.3, pp. 340–358.

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

Spoerri, T. H. (1966). "Speaking voice of the schizophrenic patient". In: *Archives of General Psychiatry* 14.6, pp. 581–585.

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

Ziv, Ido, Heli Baram, Kfir Bar, Vered Zilberstein, Samuel Itzikowitz, Eran V. Harel, and Nachum Dershowitz (2022). "Morphological characteristics of spoken language in schizophrenia patients – an exploratory study". In: *Scandinavian Journal of Psychology* 63.2, pp. 91–99. DOI: https://doi.org/10.1111/sjop.12790. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjop.12790. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sjop.12790.