Schizophrenia Bulletin vol. 49 suppl. 2 pp. S163–S171, 2023
https://doi.org/10.1093/schbul/sbac142
Advance Access publication October 28, 2022

SUPPLEMENT ARTICLE

Semantic and Acoustic Markers in Schizophrenia-Spectrum Disorders: A Combinatory Machine Learning Approach

Alban E. Voppel[*,1], **Janna N. de Boer**[1,2], **Sanne G. Brederoo**[1], **Hugo G. Schnack**[2,3], **and Iris E. C. Sommer**[1]

[1]Department of Biomedical Sciences of Cells and Systems, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; [2]Department of Psychiatry, UMCU Brain Center, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands; [3]Utrecht University, Utrecht Institute of Linguistics OTS, Utrecht, the Netherlands

[*] To whom correspondence should be addressed; Department of Biomedical Science of Cells and Systems, University Medical Center Groningen, Groningen, 9713 AV, the Netherlands; tel: +31887558672, fax: +31887555487, e-mail: A.e.voppel@umcg.nl

***Background and hypothesis*: Speech is a promising marker to aid diagnosis of schizophrenia-spectrum disorders, as it reflects symptoms like thought disorder and negative symptoms. Previous approaches made use of different domains of speech for diagnostic classification, including features like coherence (semantic) and form (acoustic). However, an examination of the added value of each domain when combined is lacking as of yet. Here, we investigate the acoustic and semantic domains separately and combined. *Study design*: Using semi-structured interviews, speech of 94 subjects with schizophrenia-spectrum disorders (SSD) and 73 healthy controls (HC) was recorded. Acoustic features were extracted using a standardized feature-set, and transcribed interviews were used to calculate semantic word similarity using word2vec. Random forest classifiers were trained for each domain. A third classifier was used to combine features from both domains; 10-fold cross-validation was used for each model. *Results*: The acoustic random forest classifier achieved 81% accuracy classifying SSD and HC, while the semantic domain classifier reached an accuracy of 80%. Joining features from the two domains, the combined classifier reached 85% accuracy, significantly improving on separate domain classifiers. For the combined classifier, top features were fragmented speech from the acoustic domain and variance of similarity from the semantic domain. *Conclusions*: Both semantic and acoustic analyses of speech achieved ~80% accuracy in classifying SSD from HC. We replicate earlier findings per domain, additionally showing that combining these features significantly improves classification performance. Feature importance and accuracy in combined classification indicate that the domains measure different, complementing aspects of speech.**

Key words: psychosis/language/speech/classification/ensemble learning

Introduction

Recently, Natural Language Processing (NLP) has become a promising marker for schizophrenia-spectrum disorders (SSD),[1–3] as it has been used successfully to predict conversion to psychosis in ultra-high-risk individuals,[4–6] distinguish healthy controls from persons/individuals with SSD[7–9] and differentiate between patients with diverse psychiatric diagnoses.[10]

Researchers have made use of different domains of speech for these investigations, including acoustic, semantic and other linguistic features with results reported around 80% accuracy.[4,6,8,11] Acoustic sources of information include pausing patterns, speech rate, intonation and percentage of spoken time.[3,12] Symptoms commonly associated with the acoustic domain include alogia (poverty of speech), blunted affect and other negative symptoms, as well as positive symptoms such as pressured speech.[13] Acoustic features of speech have also been associated with psychomotor retardation through speech motor control.[14] Semantic features include discourse coherence, semantic density and connectedness in language.[4,9,15,16] These latter approaches make use of computational models of semantic information present in language.[11] Symptoms related to these semantic methods include positive symptoms including formal thought disorder,[15,17] delusions,[18] tangentiality, and incoherence of speech.[17,19,20] Information from different domains can be overlapping, with low speech rate and low semantic density both correlating with negative symptoms.[16,21]

However, SSD is a very heterogeneous clinical picture and not all patients will have thought disorder, agitation or specific negative symptoms. Hence, not all patients will score on the same linguistic domain. We hypothesize that a combination of markers of different linguistic domains could better accommodate the high clinical heterogeneity of the phenomenology of SSD, and lead to better accuracy in classification models. Here, we investigate the relative strengths of computational models focused on different domains of speech characteristics and their combination in differentiating people with SSD from healthy controls.

When using newly developed NLP techniques in examining a heterogeneous disorder such as SSD, a multitude of features can be employed. As an example, Marmar and colleagues made use of tens of thousands of acoustic features in classifying post-traumatic stress syndrome.[22] Because of the large number of possible features, interpretation and comparability of findings is problematic.[2] Explainability of features and algorithms is a critical step towards implementing machine learning models in clinical practice.[23,24] Employing standardized and limited feature such as the GeMAPs feature set[25] in classification algorithms aids explainability and replicability, but at the potential cost of lower accuracy. Partially due to the fast developments in techniques in the semantic domain, researchers have made use of a range of semantic features, from sentence coherence using latent semantic analysis,[4,15] and word embeddings using word2vec as well as derived measures including semantic density of speech in SSD.[9,16] Recent techniques like BERT have seen usage, with models incorporating semantic context to compute subtle semantic characteristics of language.[19]

In this study, we use a limited number of speech analyses from both the acoustic and semantic domain to train two domain-separate classifiers for classification of SSD and HC. We investigate their performance and examine top features, then merge the features in a combined classifier to assess the value of using a combinatory approach. We further explore the relative strengths and weaknesses of the acoustic and the semantic domain in classifying SSD.

Methods

Participants

Speech was recorded from a semi-structured interview of 94 participants with SSD and 73 healthy controls, adding up to 167 participants. All patients were diagnosed with schizophrenia, psychosis not otherwise specified (NOS), schizophreniform or schizoaffective disorder by the treating physician, and diagnoses were confirmed using either the CASH or the MINI diagnostic interviews.[26,27] Inclusion criteria for healthy controls were the absence of a psychiatric diagnosis and history thereof, with the exception of depression or anxiety disorders in full remission. All participants were 18 years or older and native Dutch speakers. Participants were informed that the interview was analyzed for "general experiences" to prevent participants focusing on their speech or pronunciation. After completion, participants were told the true purpose of the study, to investigate their speech. Before enrollment, all participants gave written informed consent. The study was approved by the University Medical Center Utrecht ethical review board. Antipsychotic medication use was calculated as chlorpromazine equivalents in milligram per day.[28] Symptoms were assessed using the positive and negative syndrome scale (PANSS).[29]

Interview Procedure

Speech was recorded using a digital TASCAM DR-40 recording device using head-worn AKG-C544l cardioid microphones, with separate recording channels for participant and interviewer with a sampling rate of 44.1 kHz. The semi-structured interview was performed by trained interviewers; for an elaborate description of the interview methodology, see previous reports from our group.[9,21,30] Topics discussed in the interview were neutral, avoiding specific illness related topics, and participants could skip questions if they wanted. For a list of questions, see supplemental Table 1 (Table S1).

Acoustic Domain: Processing and Parameters

To remove crosstalk the following steps were taken: 1) the "annotate silences" function in PRAAT[31] was used on the interviewer's channel; 2) all resulting speech segments in which the interviewer was silent, including joint pauses, were selected on the participants channel; 3) the resulting speech segments were concatenated to a new audio file containing only segments of the participants' speech. Using openSMILE,[32] the extended GeMAPS parameter set was used to extract a total of 88 parameters at the speaker level. The parameters can be divided into 6 temporal parameters such as speech rate, 24 frequency parameters, 43 spectral parameters such as Mel-frequency cepstral coefficients, and 14 energy/amplitude parameters such as intensity. We previously published using this method, and chose this standard feature set to improve generalizability and comparison of findings across studies.[12] Moreover, the set also includes features consistently associated with psychosis in a recent meta-analysis.

Semantic Domain: Processing and Parameters

Speech was transcribed according to the CLAN-CHILDES protocol.[33] Following transcription, the text was vectorized using a 300-dimensional word2vec language model trained on a corpus of spoken Dutch.[34,35] Following previous research from our group,[9] a moving window approach was employed to calculate word-to-word similarity within windows of words sized 5-10. Within a window, the similarity of each word to each other word within that window was computed and then averaged, resulting in a single word similarity per window. The window then moved one position further, a new similarity value was computed, and this procedure was repeated until the end of the transcript, resulting in a series of word similarity values. From this series of similarity values, variance, mean, maximum and minimum of similarity between windows 5 and 10 was computed per subject, for a total of 24 semantic parameters.

Random Forest Classifier Models

To ensure comparability with earlier results and internal estimations of feature importance we chose random forest classifier algorithms.[9,12,22,36] Separate random forest classifiers were trained to assess feature accuracy per domain (i.e. acoustic and semantic). Models were trained in R, using the caret software package.[37,38] The models used 10-fold cross-validation, where 90% of the data set is used as training with a randomly chosen 10% as a testing sample, repeated ten times until all samples have served as a testing sample. 500 trees were grown, with number of features sampled per decision split the square root of the total number of features. To combine predicting features from different domains a third model was trained. The features of both the acoustic and semantic domains were used as input in this final random forest model. This model was then also cross-validated using 10-fold cross-validation.

Probability estimates for each of the trained models (acoustic, semantic and combined) were used to generate receiver operator curves (ROC) and areas under the curve (AUC). From each of the trained classifiers, predictor feature importance (Gini-importance score) was calculated, measuring how much worse the model becomes when replacing each predictor in decision trees with random data distributed according to the SSD:HC ratio. The difference between the original model performance and performance without the feature is then taken as the added value of the feature.

Statistics

Statistical analysis was performed in R.[37] Demographic characteristics were compared between the groups using ANOVAs for continuous variables, and χ2 tests were used for categorical variables. Pearson's correlational analyses were performed between continuous variables as possible confounders. To assess relative model performance significance, we used McNemar's test.[39]

Results

Participants

The participants with SSD and the HCs did not differ significantly in age or sex, see table 1. Healthy controls had received significantly more education than SSD patients,

Table 1. Demographics.

		SSD patients	HC	
Category		N = 94	N = 73	**Statistics**
Age				
Years	M (SD)	33.6 (13.4)	36.1 (15.8)	F=2.53, P= .267
Sex				
Male	n (%)	68 (72)	49 (67)	χ^2=0.533, P= .465
Years of education				
Participant	M (SD)	12.7 (2.5)	14.6 (2.2)	F= 9.248, P=.003[*]
Parental	M (SD)	12.4 (2.9)	12.1 (3.1)	F= 0.595, P=.442
Chlorpromazine dose				
milligram equivalent	M (SD)	226.2 (156.2)[a]		
Illness duration				
Years	M (SD)	4.6 (9.21)		
Diagnosis				
Psychosis NOS	n (%)	41 (44)		
Schizophrenia	n (%)	36 (38)		
Schizoaffective	n (%)	15 (16)		
Schizophreniform	n (%)	2 (2)		
PANSS				
Positive	M (SD)	11.5 (4.2)		
Negative	M (SD)	13.1 (4.6)		
General	M (SD)	26.4 (6.6)		
Total	M (SD)	51.0 (12.0)		

Legend: SSD: Schizophrenia spectrum disorder, HC: healthy control, M: mean, SD: standard deviation, NOS: Not Otherwise Specified, PANSS: Positive And Negative Syndrome Scale,

[a]: 6 subjects used antipsychotic medication for which no dosage equivalent could be calculated, thus the mean dosage calculated is of 88 subjects.

[*] denotes significant difference with $P < .05$.

Table 2. Top Informative Features of Combined Classifier.

		SSD	HC	ANOVA		Correlation with YOE
Feature name	Reflecting	Mean (sd)	Mean (sd)	*F*	*p*	Pearson's *r*
Voiced segments per sec	Average number of continuous voiced regions per second (more segments indicates more fragmented speech with short speech segments).	1.65 (0.451)	1.33 (0.212)	39.009	<.001**	−.247**
Variance 7	Variance of word connectedness, window size 7	0.0062 (0.00077)	0.0057 (0.00031)	22.126	<.001**	n.s.
SlopeV5001500_sma3nz_amean	Mean spectral Slope 0-500 Hz and 500-1500 Hz, reflecting tension	-0.0229 (0.00335)	-0.0248 (0.00301)	2.168	0.143	−.191*
Variance 10	Variance of word connectedness, window size 10	0.00622 (0.000713)	0.00571 (0.000293)	24.811	<.001**	n.s.
Variance 8	Variance of word connectedness, window size 8	0.00622 (0.000758)	0.00570 (0.000296)	23.228	<.001**	n.s.

SSD: schizophrenia-spectrum disorders; HC: healthy controls; sd: standard deviation; YOE: Years of educations; n.s.: not significant;
[*] denotes significance at $P < .05$;
[**] denotes significance at $P < .01$.

$P = .003$, which was not unexpected, as SSD commonly develops during the educational years. We therefore investigated the relation between education and the top informative classifier features, see table 2. Groups did not differ in parental educational levels.

Acoustic Classifier

The 10-fold cross-validated random forest classifier using features from the acoustic domain had an accuracy of 81%, with a sensitivity of 89%, and specificity 70%, see table 3. The AUC-ROC was 0.82 in classifying subjects with SSD from HC. The top feature as ranked by Gini importance was voiced segments per second, reflecting more fragmented speech in the SSD group, see table 2. Top 10 features are ranked in figure 1a.

Semantic Classifier

For features from the semantic domain, the 10-fold cross-validated random forest classifier reached an accuracy of 80%. Sensitivity of the model was 81%, and specificity 78%. The AUC-ROC using semantic features was 0.83, with the top feature being variance of similarity in a window of 7, indicating an increase in variance of sentence-level word similarity in SSD compared to HC, see table 2. Gini ranking of top 10 features for the semantic domain are shown in figure 1b.

Combined Classifier

The combined classifier, trained with features from both domains, reached an accuracy of 85%, with sensitivity reaching 92% and specificity 77%. The AUC-ROC reached 0.88. The top informative feature was the top acoustic feature voiced segments per second, with the second most informative feature variance of similarity in a window of 7, the highest ranked feature for the semantic classifier; for a full list of features, see figure 2.

Comparing Classifier Performance and Misclassifications

Classification results of acoustic, semantic and combinatory classifiers are shown in table 3.

Statistically comparing classification model performance showed the combined domain classifier significantly better compared to the acoustic classifier, McNemar's test $\chi^2 = 4.800$. $P = .029$, as was the combined domain classifier compared to the semantic classifier, $\chi^2 = 6.125$, $P = .013$.

To evaluate misclassifications, we compared overlap in misclassifications per modality and in the combined model. For healthy controls, 24% of subjects misclassified by the acoustic domain classifier were also misclassified by the combined classifier; the overlap in misclassification was 33% for semantic and combined classifiers. In

Table 3. Classifier Performance for Acoustic, Semantic and Combined Models.

	Accuracy	Sens. (95% CI)	Spec. (95% CI)	AUC (95% CI)
Acoustic	0.81	0.89 (0.82–0.94)	0.70 (0.59–0.79)	0.82 (0.76–0.88)
Semantic	0.80	0.81 (0.72–0.88)	**0.78** (0.67–0.86)	0.83 (0.77–0.89)
Combined	**0.85**	**0.92** (0.84–0.96)	0.77 (0.66–0.85)	**0.88** (0.83–0.93)

Bold denotes best scoring model per measure. Legend: Sens.; sensitivity, spec.: specificity, AUC: Area under the curve, CI: confidence interval.

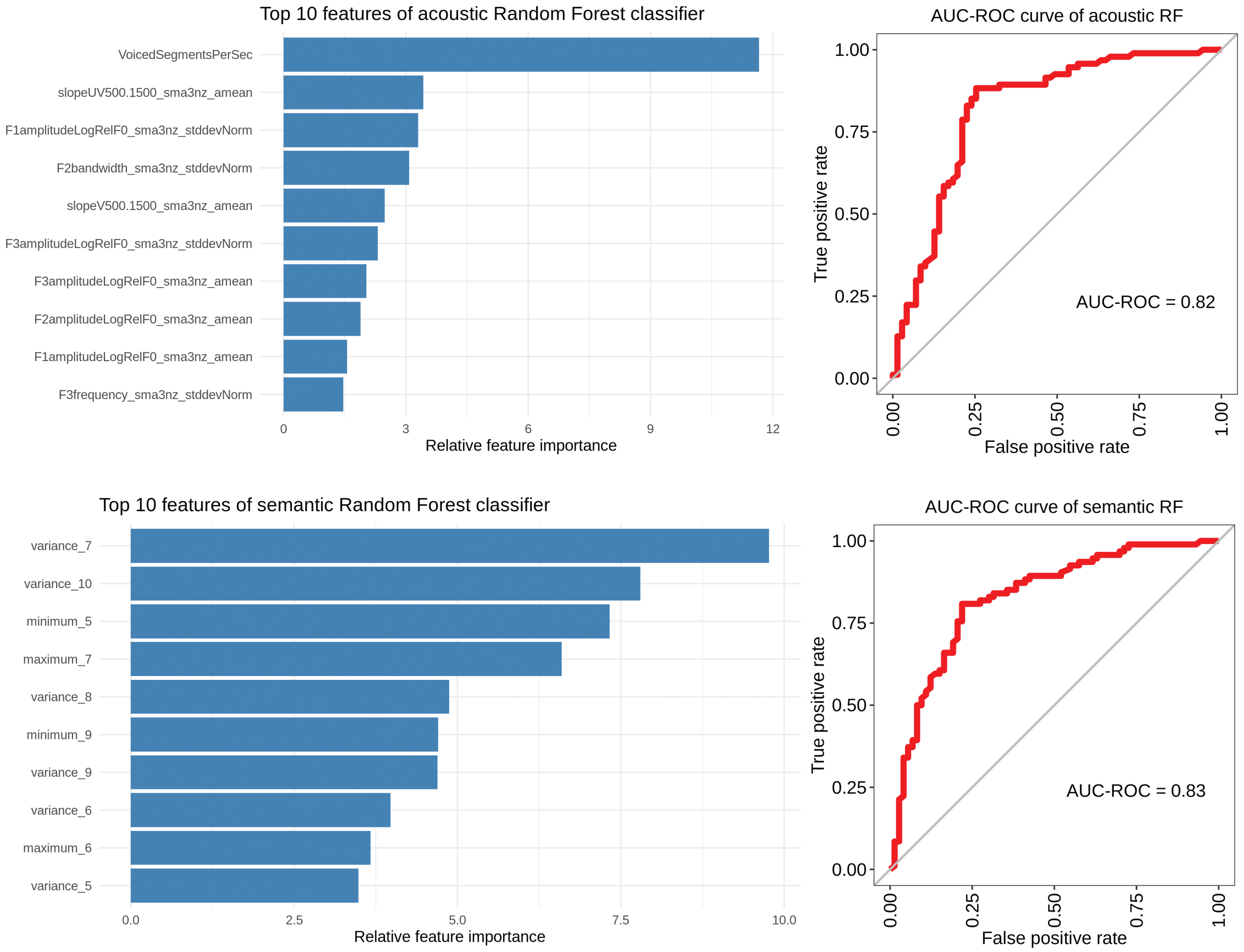

Fig. 1. Separate domain classifier performance and features. Top: Acoustic domain classifier performance. Top informative features of the acoustic classifier, left, ranked by Gini feature importance. Right, AUC-ROC curve of the acoustic classifier. **Bottom:** Semantic domain classifier performance, with top informative features left and AUC-ROC curve of semantic classifier right.

total, eight healthy controls (11%) were misclassified by all three models.

28% of SSD subjects misclassified in the acoustic classifier were also misclassified by the combined model, with a similar 28% overlap between the semantic classifier and the combined classifier SSD misclassifications. We examined positive and negative PANSS score characteristics between misclassifications of the acoustic and the semantic random forests, however these did not significantly differ (all P > .05). Only a single SSD subject was misclassified by all three models, showing the potential added sensitivity by combining different sources of information.

Discussion

This study shows that speech features of acoustic and semantic domains can be combined to improve accuracy of speech classifiers for SSD, which better captures phenomenological heterogeneity of SSD. The two highest-ranked features in the combined classifier were the two top features

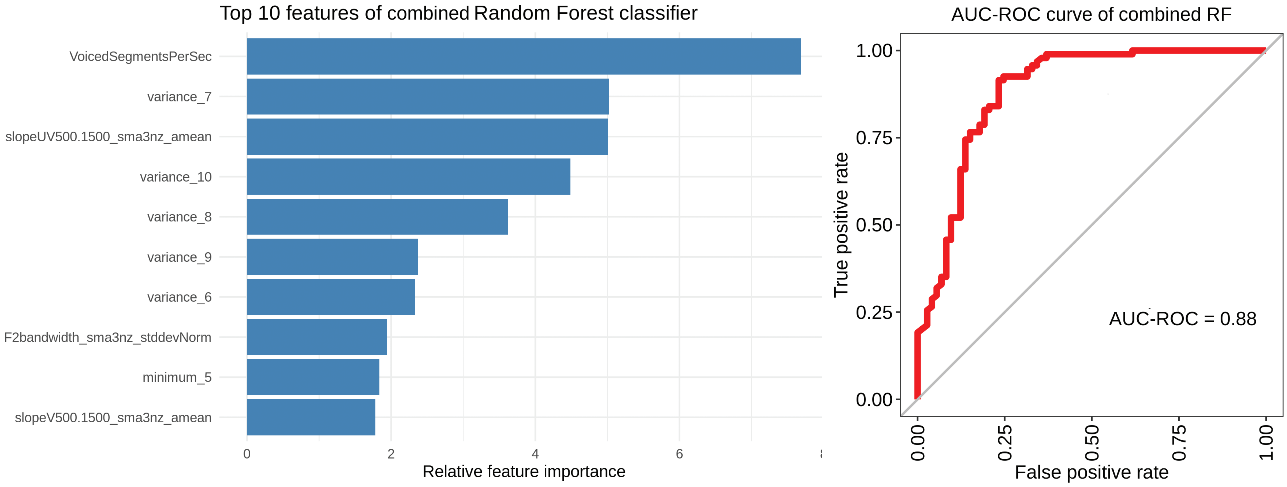

Fig. 2. Combined domains classifier Left; top features of random forest classifier, ranked by Gini coefficient. Right, AUC-ROC curve of the combined classifier.

in the domain-specific classifiers, showing the added value of combining domains to improve classifier performance. Classifiers trained on acoustic and semantic features separately were comparable in overall accuracy, with some relative strengths for sensitivity (acoustic domain) and specificity (semantic domain). The classifier combining feature domains performs significantly better than the separate models on overall performance, accuracy, as well as sensitivity.

In training separate classifiers for both the acoustic and semantic domain, we replicate previous findings showing their applicability in classifying subjects with SSD from HC.[3,4,9,12,16,40] The most informative acoustic feature, voiced segments per seconds, was supported by previous literature, as speech rate and pausing patterns are often disturbed in SSD.[3] For the semantic classifier we found that the most informative feature to be variance of word connectedness at window size seven, which was also reported previously research by our group in a smaller sample.[9] The window size of seven reflects word similarity at the sentence level (with the average length of a Dutch spoken sentence being seven words),[41] which consistently outperforms word-to-word similarity in classification algorithms.[4,11]

These findings indicate the strong presence of acoustic features in speech of SSD when compared to healthy controls. In this sample, the most informative feature was an acoustic approximation of speech rate (i.e. voices segments per second). This is an encouraging finding for future research, because acoustic features are much easier to acquire than semantic ones, as the latter require time-consuming transcriptions. Moreover, acoustic feature analyses are transferrable over languages and thus allow for crosslinguistic comparisons and the combining of data from different countries. Generalizability of speech characteristics across languages is an ongoing field of research with implications for eventual applicability of speech as a marker for SSD.[42]

Our results show that speech analyses are sensitive to subtle psychotic symptoms, since most patients in our sample were in remission and symptom scores were low. In previous work, we have shown that acoustic features can identify patients with more positive symptoms compared to those with more negative symptoms, which is likely an approximation of more acute versus more chronic symptoms, as negative symptoms are more resistant to antipsychotics.[40] We therefore expect models like these to perform even better on patients with more acute or more severe symptoms. Acoustic features might therefore be suitable to recognize relapse into psychosis. This can be an important avenue for future research, since risk of relapse after the first episode of psychosis is high even with maintenance treatment, yet predicting relapse remains challenging.[43,44]

Limitations of the Study

There was a substantial difference in years of education (YOE) between the groups. While no significant correlations were found between education and the semantic features used in this study, as expected based on previous findings,[9] we found significant correlations between years of education and the most informative acoustic features. It is thus possible that years of education explains part of the acoustic feature importance, serving as a confounder. We did not find evidence for a direct relationship between years of education and fragmented speech in the literature; however, previous research has shown that speech rate and pausing patterns differ between healthy individuals from different social backgrounds.[45,46] Although social background of course consists of more than education alone, it could be that part of this relationship is explained by education levels.

The acoustic measures chosen here, the eGeMAPS feature set, can be influenced by background noise. Subjects of both groups were interviewed in quiet rooms and were instructed not to touch the head-worn microphone after the interview started to prevent interference and hold the microphone at the same distance from the mouth. Through the crosstalk removal procedure employed, incidental background noise such as the sound of a closing door that is captured on both channels is removed; however, other sources of background noise can remain. The interviewers flagged interviews where specific events occurred in order to evaluate possible noise (i.e. in one interview a mobile phone rang; we removed the audio surrounding this event).

Furthermore, research has shown that the acoustic features have relatively low test-retest validity within subjects over time[47]; the most informative features found here, voiced segments per second and the mean slope of the spectrum between 500 and 1500 Hz have a standard error of measurement of 0.31 and 0.005, respectively. These measures indicate reliability over time through characterization of measurement error. Future research should thus take test-retest variability per measure, as well as interview characteristics such as background noise procedures into account.

Some other limitations should be mentioned too. The present study made use of cross-validation to estimate the generalizability of the models; while this is a valid approach, the usage of an independent, large validation sample is the gold standard approach to test for possible overfitting on the training set, as using cross-validation carries the risk of misestimating performance of classification models.[48,49] The sharing of datasets and code for validation using an open science approach could be invaluable for replications in an heterogeneous population, if the privacy-sensitive nature of recordings can be overcome.

In sum, our results show that, using features derived from previous findings, speech feature domains can be combined to reach greater accuracy in classifying SSD patients and healthy controls. Importantly, we here show the relative features both per domain and in a combination, allowing us to retrieve the added value of features. Adding more features allows for more complex models that better handle heterogeneous group classification, but more complex models make it harder to explain results. Similarly, various machine learning algorithms can suffer from a lack of explainability, with "black box" models being undesired in the clinical application of machine learning models.[50] Indeed, explainability of models and the importance of features therein can be a reason for choosing a slightly worse-performing classification algorithm over a more accurate one that lacks explainability.[24] Through assessing and choosing specific features, explainable and simple algorithms can be created for future clinical applications.[23]

Future research could further improve the findings presented here. While we used a relatively sparse set of semantic features and a standard feature set for acoustic assessment of speech, the addition of feature sets or approaches tailored towards specific symptoms such as impaired metaphor usage or topic prompts such as dreams could lead to improvements of results.[10,51] New techniques in the field of NLP incorporate contextual embeddings using the BERT architecture can extract more fine-grained semantic information[19,52] In addition, future research could investigate the additional value of features from other domains of language and speech, such as syntax, as these domains or features capture more of the heterogeneity present in schizophrenia-spectrum disorders. Although here we found no significant differences in symptom scores between misclassifications of the acoustic and semantic models, further comparison of for instance true versus false positives could shed more light on classifier performance. Similarly, model sensitivity analysis on (sub)groups such as SSD with specific symptom profiles, women versus men or ethnic minorities could further increase interpretability of the models as well as give information regarding their possible bias and generalizability.[53]

To lessen the burden for participants, future studies should investigate the minimal required amount of speech to make an accurate classification for each domain. Finally, while the random forest algorithm employed here ensures comparability with previous research and has a built-in measure of feature importance, other machine learning methods with differing suitability to different feature sets exist.[54] Due the fast developments in the field, a standard is as of yet absent, and the optimal algorithm might differ per dataset.

While we focused here on classifying subjects, a similar combinatory methodology can also be used for different applications such as differential diagnosis, monitoring treatment or relapse prevention. In these cases, including other specific features relevant to the application is sensible. For example, to assess comorbid depression one might include measures such as semantical sentiment analyses, while a researcher investigating the severity of cognitive impairments might want to incorporate syntactic complexity of sentences or increased pausing as a specific domain.[55]

Once a recording is available, acoustic and linguistic characteristics like syntax, semantics and sentiment analysis can be derived with no added burden for participants. The combination of reproducible, objective features spanning different domains derived from a single recording makes language and speech analysis a prime target as a marker for SSD.[56] A similar approach can be taken for other clinically informed questions such as relapse prediction or differential diagnosis by calculating features for the relevant speech characteristics linguistic features, and has been shown to be appealing to patients in the aid of their mental welfare.[57]

Concluding, we showed that acoustic and semantic features of speech can be combined to classify individuals with SSD from healthy controls with 85% accuracy. While both domains of features can be used separately, the combination of domains performs significantly better.

Supplementary Material

Supplementary material is available at https://academic.oup.com/schizophreniabulletin/.

Acknowledgements

We are grateful to all participants of the study, as well as to all research interns for their help with data collection and preparation.

Conflict of interests

The authors report no conflicts of interest.

References

1. Corcoran CM, Mittal VA, Bearden CE, *et al*. Language as a biomarker for psychosis: a natural language processing approach. *Schizophr Res.* 2020;226:158–166.
2. de Boer JN, Brederoo SG, Voppel AE, Sommer IEC. Anomalies in language as a biomarker for schizophrenia. *Curr Opin Psychiatry.* 2020;33(3):212–218. doi:10.1097/YCO.0000000000000595.
3. Parola A, Simonsen A, Bliksted V, Fusaroli R. Voice patterns in schizophrenia: a systematic review and Bayesian meta-analysis. *Schizophr Res.* 2020;216:24–40.
4. Corcoran CM, Carrillo F, Fernández-Slezak D, *et al*. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry.* 2018;17(1):67–75. doi:10.1002/wps.20491.
5. Spencer TJ, Thompson B, Oliver D, *et al*. Lower speech connectedness linked to incidence of psychosis in people at clinical high risk. *Schizophr Res.* 2021;228:493–501. doi:10.1016/j.schres.2020.09.002.
6. Bedi G, Carrillo F, Cecchi GA, *et al*. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr.* 2015;1(1):15030. doi:10.1038/npjschz.2015.30.
7. Bar K, Zilberstein V, Ziv I, Baram H, *et al*. *Semantic Characteristics of Schizophrenic Speech*. 2019:84–93.
8. Elvevåg B, Foltz PW, Rosenstein M, DeLisi LE. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J Neurolinguistics.* 2010;23(3):270–284. doi:10.1016/j.jneuroling.2009.05.002.
9. Voppel AE, de Boer J, Brederoo S, Schnack H, Sommer I. Quantified language connectedness in schizophrenia-spectrum disorders. *Psychiatry Res.* 2021;304:114–130. doi:10.1016/j.psychres.2021.114130
10. Mota NB, Furtado R, Maia PPC, Copelli M, Ribeiro S. Graph analysis of dream reports is especially informative about psychosis. *Sci Rep.* 2014;4:1–7. doi:10.1038/srep03691.
11. de Boer JN, Voppel AE, Begemann MJH, Schnack HG, Wijnen F, Sommer IEC. Clinical use of semantic space models in psychiatry and neurology: a systematic review and meta-analysis. *Neurosci Biobehav Rev.* IEC2018;93(June):85–92. doi:10.1016/j.neubiorev.2018.06.008.
12. De Boer JN, Voppel AE, Brederoo SG, *et al*. Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. *Psychol Med.* 2021;(August):1–11. doi:10.1017/S0033291721002804.
13. Alpert M, Shaw RJ, Pouget ER, Lim KO. A comparison of clinical ratings with vocal acoustic measures of flat affect and alogia. *J Psychiatr Res.* 2002;36(5):347–353. doi:10.1016/s0022-3956(02)00016-x.
14. Cannizzaro MS, Cohen H, Rappard F, Snyder PJ. Bradyphrenia and bradykinesia both contribute to altered speech in schizophrenia: a quantitative acoustic study. *Cogn Behav Neurol.* 2005;18(4):206–210.
15. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr Res.* 2007;93(*1–3*):304–316. doi:10.1016/j.schres.2007.03.001
16. Rezaii N, Walker E, Wolff P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophr.* 2019;5(1):9. doi:10.1038/s41537-019-0077-9.
17. Mackinley M, Chan J, Ke H, Dempster K, Palaniyappan L. Linguistic determinants of formal thought disorder in first episode psychosis. *Early Interv Psychiatry.* 2020;15(2):1–8. doi:10.1111/eip.12948.
18. Hinzen W, Rossello J, McKenna P. Can delusions be understood linguistically? *Cogn Neuropsychiatry.* 2016;21(4):281–299. doi:10.1080/13546805.2016.1190703.
19. Tang SX, Kriz R, Cho S, *et al*. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *npj Schizophr.* 2021;7(1):25. doi:10.1038/s41537-021-00154-3.
20. Bedi G, Carillo F, Cecchi G, *et al*. Automated analysis of disorganized communication predicts transition to psychosis among clinical high risk patients. *Neuropsychopharmacology.* 2013;38:S436–S437. doi:10.1038/npp.2013.281.
21. de Boer JN, van Hoogdalem M, Mandl RCW, *et al*. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. *npj Schizophr.* 2020;6(1):10. doi:10.1038/s41537-020-0099-3.
22. Marmar CR, Brown AD, Qian M, *et al*. Speech-based markers for posttraumatic stress disorder in US veterans. *Depress Anxiety.* 2019;36(7): 607–616. doi:10.1002/da.22890.
23. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. *PMLR*. 2019;(October):359–380. http://arxiv.org/abs/1905.05134.
24. Chandler C, Foltz PW, Elvevåg B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophr Bull.* 2020;46(1):11–14. doi:10.1093/schbul/sbz105.
25. Eyben F, Scherer KR, Schuller BW, *et al*. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Trans Affect Comput* 2016;7(2):190–202. doi:10.1109/TAFFC.2015.2457417.
26. Andreasen NC, Flaum M, Arndt S. The Comprehensive Assessment of Symptoms and History (CASH): an instrument for assessing diagnosis and psychopathology. *Arch Gen Psychiatry.* 1992;49(8):615–623.
27. Sheehan D, Janavs J, Baker R, *et al*. MINI-Mini International neuropsychiatric interview-english version 5.0. 0-DSM-IV. *J Clin Psychiatry.* 1998;59:34–57.
28. Leucht S, Samara M, Heres S, Patel MX, Woods SW, Davis JM. Dose equivalents for second-generation antipsychotics: the minimum effective dose method. *Schizophr Bull.* 2014;40(2):314–326.
29. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2):261–276. doi:10.1093/schbul/13.2.261.

30. de Boer JN, Voppel AE, Brederoo SG, Wijnen FNK, Sommer IEC. Language disturbances in schizophrenia: the relation with antipsychotic medication. *npj Schizophr.* 2020;6(1):24. doi:10.1038/s41537-020-00114-3.
31. Boersma P, Weenink DJM. *Praat: Doing Phonetics by Computer (Version 6.0.37)*. Amsterdam: Institute of Phonetic Sciences of the University of Amsterdam. 2013.
32. Eyben F, Weninger F, Gross F, Schuller B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In: *Proceedings of the 21st ACM International Conference on Multimedia*. ACM; 2013:835–838.
33. MacWhinney B. *The CHILDES Project: Tools for Analyzing Talk: Volume I: Transcription Format and Programs, Volume II: The Database*. 2000.
34. Mikolov T, Chen K, Corrado G, Dean J. *Efficient Estimation of Word Representations in Vector Space*. Arxiv. 2013:1–12. doi:10.1162/153244303322533223
35. van Eerten L. Corpus gesproken Nederlands. *Ned Taalkd.* 2007;12(3):194–215.
36. Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. *Syst Sci Control Eng.* 2014;2(1):602–609. doi:10.1080/21642583.2014.956265.
37. R Core Team, Others. *R: A Language and Environment for Statistical Computing*. 2013.
38. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw.* 2008;28(5):1–26. doi:10.18637/jss.v028.i05.
39. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* 1947;12(2):153–157. doi:10.1007/BF02295996.
40. Compton MT, Lunden A, Cleary SD, *et al.* The aprosody of schizophrenia: Computationally derived acoustic phonetic underpinnings of monotone speech. *Schizophr Res.* 2018;197:392–399. doi:10.1016/j.schres.2018.01.007.
41. Wiggers P, Rothkrantz LJM. Exploratory analysis of word use and sentence length in the spoken Dutch corpus. In: Matoušek V, Mautner P, eds. *Text, Speech and Dialogue*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007:366–373.
42. Parola A, Simonsen A, Mary J, Zhou Y, Wang H. Voice patterns as markers of schizophrenia: building a cumulative generalizable approach via cross-linguistic and meta-analysis based investigation. *Medrxiv*. 2022. doi:10.1101/2022.04.03.22273354.
43. Rubio JM, Schoretsanitis G, John M, *et al.* Psychosis relapse during treatment with long-acting injectable antipsychotics in individuals with schizophrenia-spectrum disorders: an individual participant data meta-analysis. *The Lancet Psychiatry* 2020;7(9):749–761. doi:10.1016/S2215-0366(20)30264-9.
44. Ceraso A, Lin JJ, Schneider-Thoma J, *et al.* Maintenance treatment with antipsychotic drugs for schizophrenia. *Cochrane Database Syst Rev.* 2020;2020(8). doi:10.1002/14651858.CD008016.pub3.
45. Bernstein B. Social class, linguistic codes and grammatical elements. *Lang Speech.* 1962;5(4):221–240.
46. Kendall T. *Speech Rate, Pause and Sociolinguistic Variation: Studies in Corpus Sociophonetics*. Springer; 2013.
47. Stegmann GM, Hahn S, Liss J, *et al.* Repeatability of commonly used speech and language features for clinical applications. *Digit Biomark.* 2020;4(3):109–122. doi:10.1159/000511671.
48. Flint C, Cearns M, Opel N, *et al.* Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology.* 2021;46(8):1510–1517. doi:10.1038/s41386-021-01020-7.
49. Rybner A, Trenckner Jessen E, Damsgaard Mortensen M, *et al.* Vocal markers of autism: assessing the generalizability of machine learning models. *Autism Res*. 2022; 15(6):1018–1030. doi:10.1002/aur.2721.
50. Samek W, Wiegand T, Müller K-R. *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. 2017. http://arxiv.org/abs/1708.08296.
51. Gutiérrez ED, Corlett PR, Corcoran CM, Cecchi GA. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. *EMNLP 2017 - Conf Empir Methods Nat Lang Process Proc.* 2017:2923–2930. doi:10.18653/v1/d17-1316.
52. Wouts J, de Boer J, Voppel A, Brederoo S, van Splunter S, Sommer I. belabBERT: a Dutch RoBERTa-based language model applied to psychiatric classification. *arXiv.* 2021:1–15. doi:10.48550/arXiv.2106.01091.
53. Hitczenko K, Cowan HR, Mittal VA, Goldrick M. Automated coherence measures fail to index thought disorder in individuals at risk for psychosis. *Comput Linguist Clin Psychol Improv Access, CLPsych 2021 - Proc 7th Work conjunction with NAACL 2021.* 2021:129–150. doi:10.18653/v1/2021.clpsych-1.16.
54. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. . *ACM Int Conf Proceeding Ser.* 2006;148:161–168. doi:10.1145/1143844.1143865.
55. Oomen P, de Boer JN, Brederoo SG, *et al.* Characterizing speech heterogeneity in Schizophrenia-spectrum disorders. *J Abnorm Psychol.* 2021; 131(2):172–181. doi:10.1037/abn0000736.
56. Tan EJ, Rossell SL. Questioning the status of aberrant speech patterns as psychiatric symptoms. *Br J Psychiatry.* 2020;1:2.
57. Brederoo SG, Nadema FG, Goedhart FG, *et al.* Implementation of automatic speech analysis for early detection of psychiatric symptoms: what do patients want? *J Psychiatr Res.* 2021;142(August):299–301. doi:10.1016/j.jpsychires.2021.08.019.