Prediction of Negative Symptoms of Schizophrenia from Objective Linguistic, Acoustic and Non-verbal Conversational Cues

Debsubhra Chakraborty*, Shihao Xu†, Zixu Yang‡, Victoria Chua†, Yasir Tahir§,
Justin Dauwels†, Nadia Magnenat Thalmann§, Bhing-Leet Tan‡, and Jimmy Lee‡
*Institute for Media Innovation, Interdisciplinary Graduate School, Nanyang Technological University, Singapore
†School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
‡Institute of Mental Health, Singapore
§Institute for Media Innovation, Nanyang Technological University, Singapore

***Abstract*—Speech disorders are among the salient characteristics of negative symptoms of schizophrenia. Such impairments are often exhibited through disorganized speech, inappropriate affective prosody, and poverty of speech. The current method of detecting such symptoms requires the expertise of a trained clinician, which may be prohibitive due to cost, stigma or high patient-to-clinician ratio. An objective method to extract non-verbal and verbal speech-related cues can help to automate and simplify the assessment method of severity of speech-related symptoms of schizophrenia. In this paper, a novel automated method is presented which uses speech content from schizophrenic patients to predict the clinician-assigned subjective ratings of their negative symptoms. Specifically, the interviews of 50 schizophrenia patients were recorded and features related to acoustics, linguistics and non-verbal conversation were extracted. The subjective ratings can be accurately predicted from the objective features with an accuracy of 64-82% using machine learning algorithms with leave-one-out cross-validation. Our findings support the utility of automated speech analysis to aid clinician diagnosis, monitoring and understanding of schizophrenia.**

***Keywords*-schizophrenia, negative symptoms, speech impairment, linguistics, prosody, conversation**

I. Introduction

Schizophrenia, despite its relatively low lifetime prevalence, is one of the most debilitating mental illnesses [1]. Amongst many symptoms, speech disturbance is not only considered as a key negative symptom of schizophrenia, but one of the features that indicate early onset [2]. Patients with schizophrenia often exhibit problems with syntactic complexity and semantic coherence in their production of speech. The speech and language use of patients offers valuable insight into their symptoms, trajectory of recovery and reflection of the disorders in thought, aiding their identification, assessment and monitoring [3]. The assessment and monitoring of schizophrenia has been guided and, at the same time, encumbered by the need for manual clinician diagnosis through time-consuming interviews and observations.

Substantial advances in artificial intelligence and machine learning present promising avenues to develop objective clinical tools to aid clinicians. Linguistic analysis of content generated by schizophrenic patients have become a popular mode of investigation these recent years. Text analysis programs such as the Linguistic Inquiry and Word Count (LIWC) and Diction are oft-utilized tools for linguistic analysis of spoken and written content of schizophrenic patients, ranging from autobiographical narratives [4] and written essays [5] to semi-structured and structured interviews [6], [7]. One study has reported the innovative method of using automatic conversation topic modelling to predict therapy outcomes at an accuracy of 75% [8]. Notably, these studies have found significant differences in the linguistic usage and conversational topics of schizophrenia patients, demonstrating the feasibility of using linguistic categories as features to classify between schizophrenia patients and healthy controls. Of note, majority of these studies, apart from [7], utilize manual transcriptions of spoken interviews.

Automated linguistic analysis of speech impairments related to schizophrenia also employ context-based methods like Latent Semantic Analysis (LSA) and found subtle speech differences that can distinguish schizophrenia patients from healthy controls and predict onset of psychosis in high-risk youths [9]. Other context-based linguistic analysis methods (i.e. document embedding vectors from word2vec/doc2vec) have similarly been utilized by researchers to differentiate patients of other mental illnesses such as autism spectrum disorder and healthy controls [10]. One study has used document embedding vectors as features for classification of schizophrenia patients and prediction of Negative Symptoms Assessment (NSA-16) scores [11].

Apart from linguistic analysis of speech content from schizophrenia patients, other, but fewer, studies have focused on the acoustic and non-verbal speech analysis of schizophrenia speech. Atypical voice patterns in schizophrenia are associated to clinical symptoms such as blunting of affect and may be an important indicator and contributor to the social impairments schizophrenic patients face. Rapcan et al. applied acoustic analysis to digital recordings of schizophrenic patients reading aloud and were able to differentiate patients and controls with an accuracy of 79% [12]. A recent review on acoustic patterns in schizophrenic speech found that patients exhibited reduced speech rate, pitch variability and pause duration [13], pointing towards the possibility of identifying acoustic markers of schizophrenia. A handful of studies have applied automatic non-verbal conversational analysis to interviews with schizophrenic

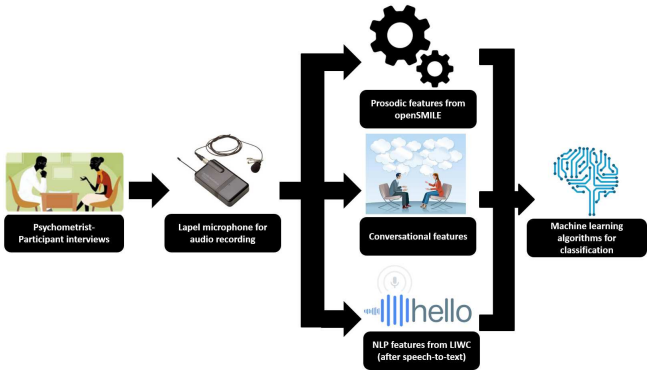

Figure 1. The facial emotions, prosodic and conversational features extraction and prediction systems.

patients and found that non-verbal conversational cues such as mutual silence, response time and natural turn-taking reliably distinguish patients and controls with an accuracy of 93%, predict NSA-16 ratings with an accuracy of 80% [14], while not reliably predict adherence to therapy [15].

Given the holistic nature of clinical assessment, speech of schizophrenic patients is manually assessed by trained clinicians according to its semantic content, syntactic coherence and conversational rapport with the interviewer [16]. Thus, combining speech, linguistic, conversational and acoustic analysis of schizophrenic patients present an important direction towards developing objective tools to aid clinician diagnosis and assessment. To that end, in this paper, we present the preliminary results of using three groups of features: (a) linguistic, (b) acoustic and (c) non-verbal conversational cues extracted from automatic transcriptions of interviews with schizophrenic patients to predict their corresponding Negative Symptoms Assessment (NSA-16) ratings. We assess machine learning algorithms (with leave-one-person-out cross-validation) and discuss the implications of our results.

II. Design of Experiment

This study was conducted in collaboration with the largest mental health institution in the region and the required ethics approval was received from the appropriate governing authorities of the region. There were 50 individuals suffering from negative symptoms of schizophrenia who participated in this study. These individuals were recruited by our clinical collaborators, and all the participants were consenting adults, and received monetary remuneration for their participation in the study. The demographics of the participants are provided in Table I.

As per the experiment design, each individual was interviewed in English by a trained psychometrician. This clinical interview was semi-structured in nature, and was audio and video recorded. The psychometrician rated the behaviour exhibited by the participant during the interview on a scale of 1-6 (1 indicating no symptoms, and 6 indicating severe symptoms of schizophrenia) on the various items of the NSA-16 [17] rating instrument. There was no role-playing involved during the interview, and participant responses were not restricted for time. The analysis of the audio recordings of these interviews has been performed for their entire duration, instead of cherry-picking a particular section of the interview. On average, the interviews lasted for about 25 minutes, and we have analysed about 21 hours of audio data. At the time of writing this manuscript, we are unaware of the existence of any other corpus containing such extensive, rich multimedia data regarding the negative symptoms of schizophrenia.

Table I
Demographics data of participants. (N = 50)

Age	Mean (years)	30.3
	Range (years)	20-46
Gender	Male	25
	Female	25
Ethnicity	Chinese	42
	Malay	5
	Indian	3
Education	University	7
	Diploma/ Vocational	27
	High School	16

III. System Overview

This section gives a short description of the audio hardware employed to record the interview of the participants, and the speech features extracted from the audio. Figure 1 gives an illustration for the overall system.

A. Sensing and recording

The audio of the participant was recorded on a portable and user-friendly H4n recorder, with lapel microphones one each for the psychometrician and the participant. The recorder was interfaced with a laptop and the audio was recorded as a 2-channel .wav file, with one channel each for the psychometrician and the participant. Since the participant was seated about 2 meters from the psychometrician, hence the cross-talk from the other channel was minimal. The audio of the participant channel was then transcribed (speech-to-text) using the Kaldi toolkit [18].

B. Linguistic Features

The text file obtained in the previous stage is processed with the latest version of Linguistic Inquiry and Word Count, or LIWC 2015 [19]. It provides a 78-dimensional feature vector with several sub-sets of words representing different

Table II
PREDICTION OF NSA-16 ITEMS FROM COMBINED LINGUISTICS, ACOUSTIC AND CONVERSATIONAL FEATURES (N = 50).

NSA-16 Item	Confusion matrix				Precision	Recall	F-score	Accuracy	Baseline Accuracy
			Predicted class						
			High	Low					
Prolonged time to respond	**True class**	High	8	8	0.89	0.50	0.64	82.00%	68.00%
		Low	1	33	0.80	0.97	0.88		
Restricted speech quantity	**True class**	High	14	7	0.78	0.67	0.72	78.00%	58.00%
		Low	4	25	0.78	0.86	0.82		
Impoverished speech content	**True class**	High	15	9	0.62	0.62	0.62	64.00%	52.00%
		Low	9	17	0.65	0.65	0.65		
Affect:Reduced modulation of intensity	**True class**	High	17	8	0.74	0.68	0.71	72.00%	50.00%
		Low	6	19	0.70	0.76	0.73		
Reduced expressive gestures	**True class**	High	10	8	0.62	0.56	0.59	72.00%	64.00%
		Low	6	26	0.76	0.81	0.79		

emotional states or characteristics, such as words related to linguistic dimensions, other grammar, and affective, social or cognitive processes. The detailed description of the evolution of LIWC 2015 and the word-subsets are available at [19]. All the word counts are normalized by the duration of the audio recording.

C. Prosodic and conversational speech features

In our previous works, we had extracted conversational and prosodic speech features from the recorded audio. The participant channel from the 2-channel .wav audio file is utilized for the computation of openSMILE prosodic features, whereas both the channels are taken into account when the conversational features for the participant are calculated. We utilized the open-Source Media Interpretation by Large feature-space Extraction, or openSMILE, toolkit [20] to calculate 988 low-level features related to emotion recognition based on the 'emobase' set [21]. The openSMILE acoustic include the following twenty-six low-level descriptors (LLD): *Intensity*, *Loudness*, *MFCC (12)*, *Pitch (F_0)*, *Probability of voicing*, F_0 *envelope*, *8 LSF (Line Spectral Frequencies)*, and *Zero-Crossing Rate*. The delta regression coefficients of the aforementioned LLDs are also computed. Over these LLDs and their delta coefficients, the following 19 measures are calculated: maximum value, minimum value, positions of the maximum and minimum values, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartiles 1, 2 and 3, and the inter-quartile ranges 1-2, 2-3 and 1-3. We have also computed 14 features associated with the dynamics of conversation between the participant and psychometrician. These non-verbal conversational cues relate to "*who* is speaking, *when*, and by *how much*" and include such features as *Number of Natural Turns*, *Interjections*, *Mutual Silence*, *Response Time* etc. [14].

IV. RESULTS

In this section, we present the results of binary classification of the relevant speech-related NSA-16 items using the combined speech features as attributes. As mentioned before, the ratings on the NSA-16 are on a scale of 1-6; however, not all of the 6 ratings are equally frequent. To overcome this problem, the 6 ratings of the NSA-16 items were re-categorized into 2 classes: Low (Class 0: ratings of 2 or below on the items, implying no observable symptom), and High (Class 1: ratings of 3 and above, implying observable symptom(s)). At each step of the leave-one-out cross-validation, we used the Kruskal-Wallis (KW) test to determine the optimal number of features; only those features with a p-value (obtained from KW test) lower than a certain threshold were retained. This optimum threshold was determined from a certain range of values, and tested on a validation set separate from the training and the left-out test set. This process was repeated for each fold of the data, hence the optimum threshold and consequently, the optimum feature sets were different (maybe only slightly) for each fold. The model was trained with Linear SVM optimized with Stochastic Gradient Descent (SGD) algorithm from Weka [22]. Table II gives the prediction results for the NSA-16 items related to emotion and speech, along with their confusion matrix, associated metrics, and accuracy. Here, the baseline accuracy is obtained as the output of a hypothetical classifier which always predicts the class-label which is more frequent of the two classes. For example, as seen from TableII, the item *Prolonged time to respond* has 34 labels as "Low", and 16 labels as "High"; so the baseline accuracy is $\frac{34}{34+16} = 68.00\%$.

V. DISCUSSION AND CONCLUSION

As can be observed from Table II, several of the NSA-16 items related to speech and emotions can be classified with high accuracy. Even the NSA-16 items related to emotion and gestures, which link indirectly to speech dysfunction, can be reliably predicted. It is often observed that individuals who speak less, also gesticulate less since speech is always accompanied by associated hand/body gestures. These symptoms are highly inter-related, and indicate towards the disruption of cognitive processes in individuals suffering from negative symptoms of schizophrenia.

Speech impairment is one of the most pronounced symptoms of schizophrenia, and is displayed through both verbal and non-verbal aspects of speech. In this paper, we utilized the combination of linguistic, acoustic and conversational signals to reliably predict a few of the subjective ratings related to speech assigned by a trained clinician. These signals can be a helpful aid in clinical practice to screen for presence and severity of negative symptoms, and even

for longitudinal monitoring of such symptoms.

VI. Acknowledgement

This study was funded by the Singapore Ministry of Health's National Medical Research Council Center Grant (NMRC/CG/004/2013) and by NITHM grant M4081187.E30. This research is also supported in part by the Being Together Centre, which in turn is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative. Moreover, this project is also funded in part by the RRIS Rehabilitation Research Grant RRG2/16009.

References

[1] J. Perälä, J. Suvisaari, S. I. Saarni, K. Kuoppasalmi, E. Isometsä, S. Pirkola, T. Partonen, A. Tuulio-Henriksson, J. Hintikka, T. Kieseppä, *et al.*, "Lifetime prevalence of psychotic and bipolar i disorders in a general population," *Archives of general psychiatry*, vol. 64, no. 1, pp. 19–28, 2007.

[2] N. C. Andreasen and W. M. Grove, "Thought, language, and communication in schizophrenia: Diagnosis and prognosis." *Schizophrenia Bulletin*, vol. 12, no. 3, p. 348, 1986.

[3] D. Gooding, S. Ott, S. Roberts, and L. Erlenmeyer-Kimling, "Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: findings from the new york high-risk project," *Psychological medicine*, vol. 43, no. 5, pp. 1003–1012, 2013.

[4] K. Hong, A. Nenkova, M. E. March, A. P. Parker, R. Verma, and C. G. Kohler, "Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls," *Psychiatry research*, vol. 225, no. 1, pp. 40–49, 2015.

[5] A. St-Hilaire, A. S. Cohen, and N. M. Docherty, "Emotion word use in the conversational speech of schizophrenia patients," *Cognitive neuropsychiatry*, vol. 13, no. 4, pp. 343–356, 2008.

[6] K. S. Minor, K. A. Bonfils, L. Luther, R. L. Firmin, M. Kukla, V. R. MacLain, B. Buck, P. H. Lysaker, and M. P. Salyers, "Lexical analysis in schizophrenia: how emotion and social word use informs our understanding of clinical presentation," *Journal of psychiatric research*, vol. 64, pp. 74–78, 2015.

[7] S. Xu, Z. Yang, D. Chakraborty, Y. Tahir, T. Maszczyk, V. Chua Yi Han, J. Dauwels, D. Thalmann, N. M. Thalmann, B.-L. Tan, and J. Lee, "Automated Lexical Analysis of Interviews with Schizophrenic Patients," in *Proceedings of the 9th International Workshop on Spoken Dialogue Systems (IWSDS)*, 2018.

[8] C. Howes, M. Purver, and R. McCabe, "Using conversation topics for predicting therapy outcomes in schizophrenia," *Biomedical informatics insights*, vol. 6, pp. BII–S11 661, 2013.

[9] C. M. Corcoran, F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. C. Javitt, C. E. Bearden, and G. A. Cecchi, "Prediction of psychosis across protocols and risk cohorts using automated language analysis," *World Psychiatry*, vol. 17, no. 1, pp. 67–75, 2018.

[10] J. Yuan, C. Holtz, T. Smith, and J. Luo, "Autism spectrum disorder detection from semi-structured and unstructured medical data," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2017, no. 1, p. 3, 2016.

[11] S. Xu, Z. Yang, D. Chakraborty, Y. Tahir, T. Maszczyk, V. Chua Yi Han, J. Dauwels, D. Thalmann, N. M. Thalmann, B.-L. Tan, and J. Lee, "Automatic verbal analysis of interviews with schizophrenic patients," Unpublished.

[12] V. Rapcan, S. DArcy, S. Yeap, N. Afzal, J. Thakore, and R. B. Reilly, "Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia," *Medical Engineering and Physics*, vol. 32, no. 9, pp. 1074–1079, 2010.

[13] A. Parola, A. Simonsen, V. Bliksted, and R. Fusaroli, "T138. acoustic patterns in schizophrenia: A systematic review and meta-analysis," *Schizophrenia Bulletin*, vol. 44, no. suppl_1, pp. S169–S169, 2018.

[14] Y. Tahir, D. Chakraborty, J. Dauwels, N. Thalmann, D. Thalmann, and J. Lee, "Non-verbal speech analysis of interviews with schizophrenic patients," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5810–5814.

[15] C. Howes, M. Purver, R. McCabe, P. G. Healey, and M. Lavelle, "Predicting adherence to treatment for schizophrenia from dialogue transcripts," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2012, pp. 79–83.

[16] B. N. Axelrod, R. S. Goldman, and L. D. Alphs, "Validation of the 16-item negative symptom assessment," *Journal of psychiatric research*, vol. 27, no. 3, pp. 253–258, 1993.

[17] N. C. Andreasen, "Negative symptoms in schizophrenia: definition and reliability," *Archives of general psychiatry*, vol. 39, no. 7, pp. 784–788, 1982.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[19] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[21] D. Chakraborty, Z. Yang, Y. Tahir, T. Maszczyk, J. Dauwels, N. M. Thalmann, J. Zheng, M. Yogeswari, N. Amirah, B.-L. Tan, and J. Lee, "Prediction of Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[22] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.