PREDICTION OF NEGATIVE SYMPTOMS OF SCHIZOPHRENIA FROM EMOTION RELATED LOW-LEVEL SPEECH SIGNALS

Debsubhra Chakraborty[†+], *Zixu Yang*[‡+], *Yasir Tahir*[†], *Tomasz Maszczyk*[*], *Justin Dauwels*[*], *Nadia Thalmann*[†], *Jianmin Zheng*[†], *Yogeswary Maniam*[‡], *Nur Amirah*[‡], *Bhing Leet Tan*[‡] *and Jimmy Lee*[‡]

[†]Institute for Media Innovation, Nanyang Technological University, Singapore
[‡]Institute of Mental Health, Singapore
[*]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

Negative symptoms of schizophrenia are often associated with the blunting of emotional affect which creates a serious impediment in the daily functioning of the patients. Affective prosody is almost always adversely impacted in such cases, and is known to exhibit itself through the low-level acoustic signals of prosody. To automate and simplify the process of assessment of severity of emotion related symptoms of schizophrenia, we utilized these low-level acoustic signals to predict the expert subjective ratings assigned by a trained psychologist during an interview with the patient. Specifically, we extract acoustic features related to emotion using the openSMILE toolkit from the audio recordings of the interviews. We analysed the interviews of 78 paid participants (52 patients and 26 healthy controls) in this study. The subjective ratings could be accurately predicted from the objective openSMILE acoustic signals with an accuracy of 61-85% using machine-learning algorithms with leave-one-out cross-validation technique. Furthermore, these objective measures can be reliably utilized to distinguish between the patient and healthy groups, as supervised learning methods can classify the two groups with 79-86% accuracy.

Index Terms— schizophrenia, affective prosody, negative symptoms assessment, openSMILE, supervised learning.

1. INTRODUCTION

Schizophrenia is a chronic and disabling mental disorder with a high probability of genetic risk inheritance [1]. The disease often develops in adolescence of an individual and adversely affects the daily functioning for the rest of her life. The presentation of schizophrenia is diverse and can be characterized broadly by positive (hallucinations and delusions), negative (apathy, blunting of affect and alogia) and cognitive (attention, memory and executive functioning) symptoms [2]. The positive symptoms are conspicuous and possess effective pharmacological treatments. However, the negative and cognitive symptoms often go undetected and neglected, and are trivialized as "laziness", although they have been consistently reported to contribute to the observed disability in schizophrenia. Furthermore, these symptoms have few to none effective drug treatments [3].

Emotional impairment is known to be one of the hallmark symptoms of schizophrenia since a long time [4]. This impairment is exhibited in both the patients' inability to express emotions through facial expressions and prosody as well as their failure to recognise displayed emotions of others [5]. In the objective to understand emotional impairment in negative symptoms of schizophrenia, blunting of facial affect has received generous attention from the scientific and medical community [6], [7]. Although emotion from speech has been studied quite extensively for healthy people [8], in comparison, speech of individuals with schizophrenia has been far less investigated [9]. Gold et al. [10] provided stimuli specifically designed to elicit emotions and associated acoustic features to a large group of patients and controls. They concluded that patients with schizophrenia are unable to identify emotion from voice due to their impairment to process low-level acoustic features, such as pitch and intensity. The authors Roux et al. [11] also concluded that explicit processing of emotional prosody is impaired in patients with schizophrenia through a vocal emotional Stroop task. Leitman et al. [12] proved the emotion recognition inability in patients were strongly related to such low-level pitch related cues as mean and variance of fundamental frequency. Association of auditory processing abnormalities with affective prosody dysfunction in schizophrenia has also been established by Jahshan et al. [13] through event-related potentials (ERPs). Occasionally, vocal and facial emotion processing deficit has been dealt together [14], through designing of crossmodal tasks. The prosody processing deficit has also been reviewed from a neuro-physiological angle in [15], [16]. Although all the methods discussed above indicate towards the same conclusion, the methods themselves are complicated, unique and non-automated. Moreover, the reverse analysis has not yet been performed, i.e., whether the low-

+ These authors contributed equally to the work.

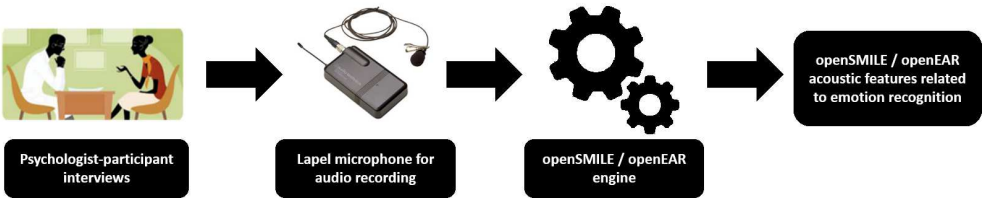

Fig. 1. The audio recording hardware and openSMILE acoustic features extraction system.

level signals, which are related to emotion, can be utilized to assess the severity of schizophrenia or to distinguish between individuals suffering from schizophrenia and healthy individuals.

We believe the low-level acoustic signals provide a strong indication of affective prosody dysfunction in schizophrenia, hence, we focus our attention to such signals in this paper. Specifically, these signals are extracted from audio recordings of participants, including schizophrenia patients and healthy control subjects, during a semi-structured clinical interview. During the interview, the patients are rated on their behaviour and social functioning on the Negative Symptoms Assessment rating instrument, one of the few purposefully developed rating instruments for the assessments of the negative symptoms of schizophrenia [17]. Currently, this method of assigning subjective ratings on a suitable assessment scale by a trained psychometrician is the only way to reliably judge the severity of negative symptoms. We wish to automate this exercise so that the subjective ratings can be predicted from the low-level prosodic cues using them as features in machine learning algorithms. We also wish to employ the same prosodic cues in the distinction of patients from healthy individuals.

This paper is organized as follows: in Section 2, we describe the experimental design, and give the demographics information regarding the participants. In Section 3, we elaborate on the hardware for audio recording and elucidate extracted the low-level acoustic signals in detail. In Section 4, we discuss our results for predicting the subjective ratings (61-85% accuracy) and classification of schizophrenic patients v/s healthy subjects (79-86%). Section 5 provides a brief discussion of the results from the previous section. Finally, in Section 6, we present our concluding remarks.

2. DESIGN OF EXPERIMENT

We collaborated with the Institute of Mental Health (IMH) in Singapore to conduct this study. Ethics approval for this study was provided by the National Healthcare Group's Domain-Specific Review Board in Singapore. The participants receive monetary remuneration for taking part in the study. All the participants are above 21 years of age and have given their written informed consent. All participants are matched for age, gender, ethnicity and educational qualifications, and are recruited by IMH. Out of the 78 total participants, there are 52 individuals with schizophrenia, grouped as *Patients*, and 26 individuals without any disorders, grouped as *Controls*. Table 1 provides the demographic information of the participants.

Table 1. Demographics data of participants.

		Patients (N = 52)	**Controls** (N = 26)
Age	Mean (years)	30.3	29.6
	Range (years)	20-46	19-47
Gender	Male	25	12
	Female	27	14
Ethnicity	Chinese	44	22
	Malay	5	3
	Indian	3	1
Education	University	7	4
	Diploma/ Vocational	28	15
	High School	17	7

Following the experiment design, each participant is assessed on a cognitive battery - the Brief Assessment of Cognition (BAC), and a semi-structured clinical interview. A trained psychometrician from IMH conducts the interviews of the participants in English. The audio, as well as the video of this interview, is recorded. The behaviour displayed by the participant during the interview is rated by the psychologist on a scale of 1-6 (1 indicating no symptoms, and 6 indicating severe symptoms) on the various items of the NSA-16 [17] rating instrument. There is no role-playing involved during the interview. As mentioned before, the interview is semi-structured, and there is no fixed time-limit for the patient responses. The audio files of these interviews have been analysed from start to finish, instead of selecting only a part of the interview. This provides a more holistic understanding of the participants' emotional behaviour in a conversation setting. On average, the interviews lasted for about 26 minutes, and we have analysed about 34 hours of recorded audio data. At this moment, we are unaware of the existence of any other dataset containing such extensive, rich multimedia data regarding the negative symptoms of schizophrenia.

3. SYSTEM OVERVIEW

In this section we briefly describe the hardware employed to record the data, and the acoustic signals extracted from such captured recordings. Fig.1 illustrates the audio recording and openSMILE acoustic features extraction system.

Table 2. Prediction of NSA-16 Items from openSMILE Audio Features for Individuals with Schizophrenia (N = 52).

NSA-16 Item	Confusion matrix				Precision	Recall	F-score	Accuracy	Baseline Accuracy	Algorithm	Feature selection
			Predicted class								
			High	Low							
Prolonged time to respond	**True class**	High	10	7	0.83	0.59	0.69	82.69%	67.31%	*k*NN	F-score
		Low	2	33	0.83	0.94	0.88				
Restricted speech quantity	**True class**	High	17	4	0.77	0.81	0.79	82.69%	59.61%	Gaussian SVM	PCA
		Low	5	26	0.87	0.84	0.85				
Impoverished speech content	**True class**	High	20	5	0.80	0.80	0.80	80.77%	51.92%	Linear SVM	Linear SVM
		Low	5	22	0.81	0.81	0.81				
Emotion reduced range	**True class**	High	16	12	0.67	0.57	0.62	61.54%	53.85%	*k*NN	Random Forest
		Low	8	16	0.57	0.67	0.62				
Affect:Reduced modulation of intensity	**True class**	High	21	5	0.78	0.81	0.79	78.85%	50.00%	Adaboosted Decision Trees	Decision Trees
		Low	6	20	0.80	0.77	0.78				
Reduced expressive gestures	**True class**	High	14	4	0.78	0.78	0.78	84.62%	65.38%	Adaboosted Decision Trees	χ^2
		Low	4	30	0.88	0.88	0.88				

3.1. Sensing and Recording

We employed easy-to-use portable equipment for recording conversations; it consisted of lapel microphones for the participant and the psychologist, and an audio H4N recorder that allowed multiple microphones to be interfaced with a laptop. The audio data was recorded as a single 2-channel audio .wav file, with one channel for each of the psychologist and the participant. The participant and the psychologist were seated at a distance of about 2 meters, hence the cross-talk in the channels were minimal. The openSMILE features were extracted only from the participant audio.

3.2. Audio features for emotion

OpenSMILE, or open-Source Media Interpretation by Large feature-space Extraction, toolkit is a modular and flexible audio feature extractor for signal-processing applications [18]. We concentrate only on the acoustic signals which are related to emotion recognition, and hence our acoustic features are based on the openSMILE 'emobase' set consisting of 988 features for emotion recognition. Following twenty-six low-level descriptors (LLD) are calculated from the audio for this set: *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. Both the LLDs and their delta coefficients are smoothed by a moving average filter with window size 3. Over these LLDs and their delta coefficients, the following nineteen 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. Consequently, the 19 measures computed over 26 LLDs and their 26 delta coefficients yield ($19 \times 26 \times 2$) 988 acoustic features.

4. RESULTS

In this section, we present the results of classification using the openSMILE audio signals for emotion. First, we present the binary classification results for a few of the relevant NSA-16 items, using the openSMILE signals as features for supervised pattern recognition classifiers. Further, we also provide the prediction results of the classification of participants into *Patient* and *Control* groups, based on the openSMILE signals only as features, as well as for the combination of openSMILE signals with body movement signals as a larger feature-set.

4.1. NSA-16 items prediction

We applied the openSMILE acoustic features for emotion recognition as objective features to predict subjective NSA-16 ratings related to emotional behaviour of the individuals suffering from negative symptoms of schizophrenia (N = 52). As mentioned before, the ratings on the NSA-16 are on a scale of 1-6; however, the distribution of the ratings is highly irregular, i.e, not all of the 6 ratings are equally frequent. To overcome this problem, the 6 ratings of the NSA-16 items are re-categorized into 2 classes: Unobservable (Class 0: ratings of 2 or below on the items, implying no observable symptom), and Observable (Class 1: ratings of 3 and above, implying observable symptom(s)).

We trained multiple machine learning classifiers with the acoustic signals as features and the class labels as targets, with leave-one-out cross-validation. The following classifiers were tested: *k* Nearest Neighbours, Decision Trees, Random Forest, Adaboosting with Decision Trees, SVM with Gaussian kernel, SVM with linear kernel, Naive Bayes, and Multilayer Perceptron. Since the openSMILE feature-set contains a total 988 features, some of the features may be irrelevant for the classification objective. To reduce the detrimental effects of confusing features on the performance of the classifier, a rigorous feature-selection method was applied. For each step of leave-one-out cross-validation, the features of the training set were ranked using one of the following techniques: F-score (ANOVA), χ^2, Mutual Information, Pearson correlation, Principal Components, linear SVM, Decision Trees, and Random Forests. Subsequently, the optimal number of features were selected according to the previous ranking methods using forward feature selection and 5-fold nested

cross-validation for each step of the main leave-one-out cross-validation. Table 2 gives the prediction results for the NSA-16 items related to emotion and speech, along with their confusion matrix, associated statistics, and accuracy for the best performing classifier and feature-selection combination. Here the baseline accuracy indicates the output of a classifier which predicts all the instances of the set as the majority class.

Table 3. Patients v/s Controls classification with openSMILE Speech Signals of Emotion.

Confusion matrix				Precision	Recall	F-score	Accuracy	Baseline Accuracy	Algorithm	Feature selection
		Predicted class								
		0	1							
True class	0	19	7	0.68	0.73	0.70	79.49%	66.67%	Linear SVM	PCA
	1	9	43	0.86	0.83	0.84				

4.2. Classification of participants

Next, we utilized the objective openSMILE acoustic signals as features for binary classification in order to distinguish between the Controls (Class 0, N = 26) and Patients (Class 1, N = 52) groups. The participant groups were given as target labels, and leave-one-out cross-validation was performed to calculate the accuracy of prediction. Feature-selection was applied as described in Section 4.1. Table 3 lists the classification accuracy, confusion matrix and associated metrics for the best classifier.

Table 4. Patients v/s Controls classification with openSMILE Speech Signals and Body Movement Signals.

Confusion matrix				Precision	Recall	F-score	Accuracy	Baseline Accuracy	Algorithm	Feature selection
		Predicted class								
		0	1							
True class	0	19	4	0.79	0.83	0.81	86.36%	65.15%	Linear SVM	F-score
	1	5	38	0.90	0.88	0.89				

In one of our earlier studies [19], we extracted 14 body movement signals based on the skeleton recorded by the Microsoft Kinect depth camera. Such signals, related to the speed and acceleration of movement of the head and upper limbs, were also utilized to distinguish between the *Controls* and *Patients* with a reasonable accuracy. These movement signals are appended to the current acoustic signals set, since it has been well-documented that emotions are also expressed through gestures and body movements [20]. We utilized this extended set to perform classification of participants again into *Controls* (Class 0, N = 23) and *Patients* (Class 1, N = 43) groups with leave-one-out cross-validation and feature-selection. For some participants, movement signals are unavailable, due to errors in video recording or lack of consent of participant to be video-recorded. Hence, the number of participants for this classification is smaller compared to the classification with only acoustic features. The results of the classification and associated metrics for the best-performing classifier are given in Table 4.

5. DISCUSSION

As can be observed from Table 2, several of the NSA-16 items can be classified with high accuracy. Even the NSA-16 items related to speech and gestures, which are indirectly linked to emotional impairment, can be reliably predicted. These symptoms are highly inter-related, and are often found to cohabit and co-exhibit in patients suffering from negative symptoms. Another inference which can be drawn from the Tables 3 and 4 is that blunting of affect results in a muted display of both prosody and associated gestures and movements, yielding high classification accuracy.

6. CONCLUSION

Schizophrenia is a chronic mental disorder and has a considerable impact on the lives of millions of people worldwide. Blunting of affect is one of the most salient negative symptoms of schizophrenia, and is commonly manifested through deficiency in affective prosody processing. In this paper, we utilized the low-level acoustic prosodic signals to reliably predict quite a few of the expert subjective ratings related to emotion assigned by a trained clinician. We also demonstrated that these prosodic signals alone, or in combination with movement signals, could also be utilized to distinguish between individuals with schizophrenia from healthy individuals with a high accuracy. These prosodic 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. In the future, we wish to utilize the video modality of our recordings to an even greater extent, by combining the analysis of facial expressions with prosody for a more holistic understanding of emotional blunting in schizophrenia.

7. ACKNOWLEDGEMENT

This study was funded by the Singapore Ministry of Health's National Medical Research Council Center Grant awarded to the Institute of Mental Health Singapore (NMRC/CG/004/2013) and by NITHM grant M4081187.E30. This research is also supported in part by the Being Together Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at Chapel Hill. The Being Together Centre is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative. Moreover, this project is funded in part by the RRIS Rehabilitation Research Grant RRG2/16009. The authors would also like to acknowledge the Interdisciplinary Graduate School and Nanyang Technological University for their support of this research.

8. REFERENCES

[1] Alastair G Cardno, E Jane Marshall, Bina Coid, Alison M Macdonald, Tracy R Ribchester, Nadia J Davies, Piero Venturi, Lisa A Jones, Shon W Lewis, Pak C Sham, et al., "Heritability estimates for psychotic disorders: the maudsley twin psychosis series," *Archives of general psychiatry*, vol. 56, no. 2, pp. 162–168, 1999.

[2] Caroline Demily and Nicolas Franck, "Cognitive remediation: a promising tool for the treatment of schizophrenia," 2008.

[3] Brendan P Murphy, Young-Chul Chung, Tae-Won Park, and Patrick D McGorry, "Pharmacological treatment of primary negative symptoms in schizophrenia: a systematic review," *Schizophrenia research*, vol. 88, no. 1, pp. 5–25, 2006.

[4] Eugen Bleuler, "Dementia praecox or the group of schizophrenias.," 1950.

[5] Jane Edwards, Henry J Jackson, and Philippa E Pattison, "Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review," *Clinical psychology review*, vol. 22, no. 6, pp. 789–832, 2002.

[6] Christian G Kohler, Jeffrey B Walker, Elizabeth A Martin, Kristin M Healey, and Paul J Moberg, "Facial emotion perception in schizophrenia: a meta-analytic review," *Schizophrenia bulletin*, vol. 36, no. 5, pp. 1009–1019, 2009.

[7] Manas K Mandal, Rakesh Pandey, and Akhouri B Prasad, "Facial expressions of emotions and schizophrenia: A review.," *Schizophrenia bulletin*, vol. 24, no. 3, pp. 399, 1998.

[8] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[9] Marjolijn Hoekert, René S Kahn, Marieke Pijnenborg, and André Aleman, "Impaired recognition and expression of emotional prosody in schizophrenia: review and meta-analysis," *Schizophrenia research*, vol. 96, no. 1, pp. 135–145, 2007.

[10] Rinat Gold, Pamela Butler, Nadine Revheim, David I Leitman, John A Hansen, Ruben C Gur, Joshua T Kantrowitz, Petri Laukka, Patrik N Juslin, Gail S Silipo, et al., "Auditory emotion recognition impairments in schizophrenia: relationship to acoustic features and cognition," *American Journal of Psychiatry*, vol. 169, no. 4, pp. 424–432, 2012.

[11] P Roux, A Christophe, and C Passerieux, "The emotional paradox: dissociation between explicit and implicit processing of emotional prosody in schizophrenia," *Neuropsychologia*, vol. 48, no. 12, pp. 3642–3649, 2010.

[12] David I Leitman, Petri Laukka, Patrik N Juslin, Erica Saccente, Pamela Butler, and Daniel C Javitt, "Getting the cue: sensory contributions to auditory emotion recognition impairments in schizophrenia," *Schizophrenia bulletin*, vol. 36, no. 3, pp. 545–556, 2008.

[13] Carol Jahshan, Jonathan K Wynn, and Michael F Green, "Relationship between auditory processing and affective prosody in schizophrenia," *Schizophrenia research*, vol. 143, no. 2, pp. 348–353, 2013.

[14] C Mangelinckx, JB Belge, P Maurage, and E Constant, "Impaired facial and vocal emotion decoding in schizophrenia is underpinned by basic perceptivo-motor deficits," *Cognitive Neuropsychiatry*, pp. 1–7, 2017.

[15] David I Leitman, Daniel H Wolf, Petri Laukka, J Daniel Ragland, Jeffrey N Valdez, Bruce I Turetsky, Raquel E Gur, and Ruben C Gur, "Not pitch perfect: sensory contributions to affective communication impairment in schizophrenia," *Biological psychiatry*, vol. 70, no. 7, pp. 611–618, 2011.

[16] Daniel C Javitt and Robert A Sweet, "Auditory dysfunction in schizophrenia: integrating clinical and basic features," *Nature Reviews Neuroscience*, vol. 16, no. 9, pp. 535–550, 2015.

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

[18] Florian Eyben, Martin Wöllmer, and Björn 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.

[19] Debsubhra Chakraborty, Yasir Tahir, Zixu Yang, Tomasz Maszczyk, Justin Dauwels, Daniel Thalmann, Nadia Magnenat Thalmann, Bhing-Leet Tan, and Jimmy Lee, "Assessment and prediction of negative symptoms of schizophrenia from rgb+ d movement signals," 2017.

[20] Pierre Feyereisen and Jacques-Dominique De Lannoy, *Gestures and speech: Psychological investigations*, Cambridge University Press, 1991.