Analyzing acoustic and prosodic fluctuations in free speech to predict psychosis onset in high-risk youths

Carla Agurto, *Senior Member, IEEE,* Mary Pietrowicz, *Member, IEEE*, Raquel Norel, Elif K. Eyigoz, Emma Stanislawski, Guillermo Cecchi, Cheryl Corcoran

***Abstract*— The diagnosis and treatment of psychiatric disorders depends on the analysis of behavior through language by a clinical specialist. This analysis is subjective in nature and could benefit from automated, objective acoustic and linguistic processing methods. This integrated approach would convey a richer representation of patient speech, particularly for expression of emotion. In this work, we explore the potential of acoustic and prosodic metrics to infer clinical variables and predict psychosis, a condition which produces measurable derailment and tangentiality in patient language. To that purpose, we analyzed the recordings of 32 young patients at high risk of developing clinical psychosis. The subjects were evaluated using the Structured Interview for Prodromal Syndromes/Scale of Prodromal Symptoms (SIPS/SOPS) criteria. To analyze the recordings, we examined the variation of different acoustic and prosodic metrics across time. This preliminary analysis shows that these features can infer negative symptom severity ratings (i.e., SIPS-Btotal), obtaining a Pearson correlation of 0.77 for all the subjects after cross-validated evaluation. In addition, these features can predict development of psychosis with high accuracy above 90%, outperforming classification using clinical variables only. This improved predictive power ultimately can help provide early treatment and improve quality of life for those at risk for developing psychosis.**

I. Introduction

Psychiatry has historically relied on observation for diagnosis and assessment of mental illness. In this process, the assessment of speech plays a very important role, but it is subjective in nature. The absence of objective clinical tests in the field can potentially be solved through a more systematic analysis of speech by means of natural language processing. Speech provides the opportunity to evaluate any mental condition, yet it may be even more important for the evaluation of psychosis.

Psychosis is usually associated with schizophrenia, but it can be caused by other mental illnesses such as depression, bipolar disorder, dementia and borderline personality disorder [1]. Psychosis is characterized by disorganized thought and irrational perception. The typical onset is late adolescence/young adulthood and it emerges earlier in males than females. Psychosis can have devastating functional implications and can lead to social exclusion [2][3][4]. Early detection of subjects at clinical high risk (CHR) of psychosis is key to avoiding these difficult repercussions. CHR has acquired a specific diagnostic category in the DSM-5 [5] in recent decades, and researchers have focused on studying people who belong to this group.

Natural language processing opens the door to an automated, objective assessment of mental disorders [6], [7]. In particular, previous work [8], [9] has demonstrated the potential of the semantic and syntactic domains of speech to predict psychosis onset. Previous research in acoustic and prosodic features of psychosis has been mostly focused on schizophrenia. Some of the relevant findings of this research are the detection of impaired intonation contours in individuals that present negative symptoms; their increased pause time [10], and reduced vowel space [11], [12]; and the differences in Mel-frequency cepstral coefficients (MFCC) between healthy controls and schizophrenia patients. With respect to psychosis, researchers have found that prosody is affected with an increase number of pauses and many hesitations than in normal speech, and first episode psychotic patients displaying significant linguistic and emotional prosody deficits compared to healthy controls [13]. The lack of automated acoustic analysis in this field provides the motivation for the present study, which aims to explore whether acoustic abnormalities can be detected in prodromal patients and, if so, whether these can be used to predict future psychosis onset.

II. Methods

A. Participants

The study was approved by the Institutional Review Board at the New York State Psychiatric Institute at Columbia University. A total of 34 subjects (ages 14 to 27) at CHR of psychosis were recruited. Adults provided written informed consent; subjects under 18 provided written assent, with consent provided by a parent. All CHR subjects in this study met criteria for the attenuated positive symptom syndrome. Exclusion criteria included history of threshold psychosis or Axis I psychotic disorder, risk of harm to self or others, and any major medical or neurological disorder. Two subjects were removed from our analysis since they had incomplete audio recordings. From the 32 subjects, five developed psychosis (CHR+) within a period of 3 to 16 months according to the SIPS/SOPS criteria, and 29 did not (CHR-). Table 1 shows the demographic information of the subjects.

B. Clinical variables

CHR subjects present attenuated psychotic symptoms which are measured using the Structured Interview for

C. Agurto*, M. Pietrowicz, R. Norel, E.K. Eyigoz, G.A. Cecchi are with Computational Biology Center, T.J. Watson IBM Research Laboratory, New York (*e-mail: capaagri@gmail.com).

E. Stanislawski is with Department of Psychiatry, Icahn School of Medicine at Mount Sinai.

C. Corcoran is with Department of Psychiatry, Icahn School of Medicine at Mount Sinai and Bronx VA MIRECC.

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

Prodromal Syndromes/Scale of Prodromal Symptoms (SIPS/SOPS). This scale is subdivided into 4 categories: positive, negative, disorganized, and general symptoms. In addition, it incorporates the Global Assessment of Functioning (GAF) scale [5], which is used to estimate the severity of a mental condition. A more comprehensive discussion of SIPS/SOPS can be found elsewhere [14]. Trained master-level research assistants administered the SIPS/SOPS, with clinical ratings achieved by expert consensus at baseline and follow-up visits. Baseline assessment only is used in this study.

Table 1: Demographic information of the participants in this study.

	CHR+	CHR-
Age	22.2 (3.4)	21.7 (3.6)
Gender (%male)	80%	59%
Race (% Caucasian)	40%	34%
Medication	20%	26%
Months to develop psychosis	8.6 (4.9)	--

Abbreviations: CHR+, clinical high-risk participants who transitioned to psychosis during follow-up; CHR-, clinical high-risk participants who did not transition to psychosis during follow-up.

C. Protocol

Oral interviews were recorded in the Center of Prevention Evaluation (COPE) between 2007 and 2012. Interviews had a duration of approximately 1 hour and were conducted by experts in qualitative research methods. Prompts queried to describe impact of life changes experienced, and their expectations for the future. All of the participants were fluent in English.

III. Experiments

A. Pre-processing

The interviewers' speech was not considered in the analysis, although their participation was minimal. After the interviewer speech was removed, the recording was divided in clips of 2-minute duration and acoustic and prosody features were calculated as specified in section III.B. Then, we calculated the standard deviation of all the data points to measure the fluctuations of the voice features along the interview time.

B. Feature Extraction

Below we summarize the features that were used to characterize CHR subjects based on previous findings in schizophrenia and psychosis.

Pitch variation

Given that one of the most well-known aspects of negative symptoms is the flattened intonation, we extract features to capture this. Intonation is primarily a matter of variation in the pitch level of the voice. Therefore, the first type of feature we explored is pitch information. To do so, we extract pitch using the autocorrelation method in the Praat toolkit [15] using 40-ms Hanning-filtered windows, and computed 6 statistical descriptors to characterize the distribution, including median, interquartile range from 25th to 75th percentiles (IQR), 5th (pct5) and 95th (pct95) percentiles and 3rd and 4th moments (skewness and kurtosis).To complement these features, we also calculate the mean and standard deviation of the glottal pulse period (pitch period length).

Spectral characterization

MFCC scores have been used to characterize schizophrenia patients from healthy controls. Previous research [16] found that some coefficients were significantly lower in the patient group than in the control group. To perform a thorough analysis, we computed 13 MFCCs coefficients across the entire durations of the speech segments as well as for the voiced and unvoiced segments of speech. In addition, from the long-term average spectra, we extracted spectral slope, median, IQR and maximum energy, and the frequency where the maximum energy value is located. To complement these spectral features, we use the library PyAudio to compute a set of 21 different spectral and chroma (pitch class profile) features [17].

Vowel space

Researchers have found that speakers with schizophrenia displayed an 11% reduction in vowel space (i.e., a reduction of tongue height captured by first formant-F1) [11], [12]. In addition, vowel space features have also been found to be informative for predicting emotional states (e.g. anger) [18], [19]. This would render vowel space features to be informative (a priori) to characterize some of the negative symptoms in the SIPS/SOPS categories. For this reason, we extract several features from the vowels using the following steps. We isolated the vowels and computed the values of formants 1, 2 and 3 and their bandwidths using [15]. For all of their distributions, we computed the 6 statistical descriptors discussed before. Additionally, we characterized the vowel space by measuring the total area, a-i-u triangle area, centroid, and orientation (slope of F2/F1).

Voice quality

Voice quality in terms of jitter (fluctuations in pitch) and shimmer (fluctuations in volume) has been studied and found to be correlated with clinically-rated hostility, a characteristic of mental illness [20] which is present in 40% of acute psychotic cases. For this reason, we incorporate these features in our analysis. To quantify these changes, we extract jitter (local absolute value and ppq5) and shimmer (local absolute and apq5) values. In addition, we computed the fraction of locally unvoiced frames and the total duration of the breaks between the voiced parts of the signal, divided by the total duration of the analyzed part of the signal. To complement these features, we also calculate the following noise measurements: mean autocorrelation, the harmonics to noise ratio (HNR), and noise to harmonics ratio (NHR).

Changes in rhythm

When the natural flow of speech is disturbed, it usually leads to a deviation in rhythm structure. This can lead to

difficulties in producing speech sounds in correct sequence, and problems finding the correct words (increasing the duration of pauses). These patterns are usually observed in individuals with psychosis. Therefore, to characterize changes in rhythm, we use the information contained in the pauses and syllables uttered by the speaker. First, we detect the syllables automatically using the method described in [21]. Once they are located, we can build a distribution of timing of syllables. In addition, we also identified the pauses using [15] by applying a threshold of -25dB and a minimum duration of 100ms in the recording. Since the use of voiced and unvoiced parts in the speech may also affect the rhythm, we also identified them in the analysis and measured their duration. From the obtained distributions for each of the categories, we compute 8 descriptors (median, IQR, pct5, pct95, skewness, kurtosis, total, and number of appearances). Besides estimating those features, we compute the speech rate (number of syllables over the total duration of the recording), articulation rate (number of syllables over the total recording time after pauses were removed), voiced/unvoiced ratio, and voiced and unvoiced percentages in the speech.

C. Experimental design

1. Experiment 1: Can we use the proposed features to infer any of the SIPS/SOPS scores?

First, we wanted to discover whether the acoustic and prosodic information of the speech in CHR subjects has the potential to provide an automated assessment of psychiatric disorders by inferring the scores obtained in the SIPS/SOPS test.

Since we are covering a broad spectrum of acoustic and prosodic features, we use LASSO linear regression to infer the clinical variables as it does both parameter shrinkage and variable selection automatically, which helps to reduce over-fitting. The validation consists of a nested 10-fold cross validation, in which we standardize the features (mean = 0 and standard deviation = 1), estimate the optimal regularization parameters from a set of values ([0.01, 0.1, 1, 10, 100]), and calculate the performance of the model. To measure the prediction performance, we use Pearson correlation.

In addition, we replicated the analysis based on gender cohorts separately, since studies have found differences between genders. For example, researchers in [22] indicate that while males initially show negative/cognitive symptoms, females present (sub-threshold) positive and affective symptoms. This also agrees with [23], where the authors state that males at high risk of psychosis displayed more pronounced negative symptoms, higher rates of past substance abuse disorders and higher deficits in social functioning than females.

2. Experiment 2: Can the proposed features predict which of the clinical high-risk participants will transition to psychosis during follow-up assessment?

Second, we wanted to evaluate if the proposed features could detect whether CHR subjects will transition to psychosis (CHR+). From our previous research in speech content [8], [9], we learned that subjects can present different symptoms (not unique for all CHR+) that lead to psychosis, which may require complex, nonlinear classifiers. Therefore, we used radial basis functional kernel SVM, which is used to classify data that is not linearly separable. However, since we are looking at fluctuations and use a rich of set of features, we also included logistic regression, a linear classifier, with l1-norm penalty. The same cross validation approach explained in Experiment 1 was applied for this task. For logistic regression we optimized the regularization parameter from these values: [0.01, 0.1, 1, 10, 100, 1000]. In the case of RBF-SVM, we optimized the regularization parameter and the inverse of the standard deviation of the RBF kernel (gamma) from these set of values: [0.01, 0.1, 1, 10, 100], and [0.01, 0.03, 0.06, 0.16, 0.40, 1, 1.78, 3.16, 5.62, 10]. Performance was measured in terms of accuracy and AUC. In addition, since it was performed for experiment 1, we also analyzed the data independently for each gender category group.

Table 2: Pearson Correlation values of the cross-validated approach

SIPS/SOPS scores	All (32)	Female (11)	Male (21)
A.1 Delusional ideas	0	0	0
A.2 Persecutory ideas	0.15	0.36	0
A.3 Grandiosity	0.27	0	0.17
A.4 Hallucinations	0.36	0.67	0
A.5 Conc. Disorganization	0	0	0
A total – Positive Symptoms	**0**	**0.84**	**0**
B.1 Social isolation	0.43	0.41	0.57
B.2 Avolition	0.39	0.72	0.52
B.3 Dec. Expression of Emotion	0.50	0.72	0.45
B.4 Dec. Experience Emotion or self	0.24	0.27	0.47
B.5 Dec. Idea. Richness	0.37	0	0.42
B.6 Det. In Role functioning	0.60	0.45	0.27
B total – Negative Symptoms	**0.77**	**0.59**	**0.55**
C.1 Odd Behavior	0.05	0.23	0.43
C.2 Bizarre thinking	0	0.44	0.24
C.3 Trouble focusing	0.18	0.79	0
C.4 Personal hygiene	0	0	0
C total - Disorganization	**0.18**	**0.40**	**0**
D.1 Sleep disturbance	0.41	0.44	0
D.2 Dysphoric mood	0	0	0.09
D.3 Motor disturbances	0.03	0	0
D.4 Impaired tolerance-stress	0	0	0
D total - General	**0.18**	**0**	**0**
Global Assessment of Functioning (GAF)	**0.56**	**0.71**	**0.23**

IV. Results and Discussion

A. Experiment 1: Infer SIPS/SOPS scores

Table 2 presents the results of inferring the clinical variables for all the SIPS/SOPS main categories as well as its sub-categories. It can be observed that, overall, our features can infer negative symptoms (SIPS/SOPS-B) with high correlation regardless of gender. This is important since it

has been found that subjects that are CHR present profound negative symptoms [24]. These results are encouraging since we can extend the work to not only assess subjects at CHR, but also identify them from the population. We also observe that there are categories such as odd behavior and bizarre thinking (C.1 and C.2) for which our model cannot infer these scores for all the subjects. However, when the subjects are divided by gender, we obtain correlation scores above 0.40. We speculate that these symptoms are very gender specific, thus requiring gender-specific models. In particular, we observe that we can infer the SIPS-A total score with Pearson value of 0.84 for females. However, our features can only infer two of the 6 sub-scores for females. Posterior analysis showed that (A.2+A.4) and Atotal have a correlation of 0.77. This will be the main reason why we obtain high prediction for A total.

In the case of GAF scores, the prediction is better for females than males. Our data shows that the mean (std) GAF scores for males and females are 43.9 (8.4) and 50.9 (8.4) respectively. This means that males in this study have more psychological problems than females that may not be captured by our features. Finally, we note that scores like delusional ideas (A.1), which we were not able to infer, may be more related to the content of the speech rather than the acoustic components. Therefore, we believe by combining the content and acoustic speech features, we can infer more accurate and a broader range of symptoms of the SIPS/SOPS.

B. Experiment 2: Predict transitioned into psychosis

Tables 3 shows the performance of our method in predicting future conversion to psychosis. We observed that for this task, the classification using all the subjects is not good. The selected classifiers only achieved chance classification accuracy. However, when the features are corrected by gender, the AUC score improved 25%. This means that the path to prediction which subjects will develop psychosis is gender specific. The results improved even further if we only create a model for males obtaining AUC and accuracy scores of 0.99 and 0.95 with logistic regression classifier. For this particular case, we found that the top 5 relevant features were kurtosis and skewness of B1 (bandwidth of F1), pct95 of F2, IQR Pitch, and MFCC#9. It has been found that vowel space features and variation in intonations (pitch) are informative for detecting depression [25], [26], and in turn may affect social functioning, a key component that predicts conversion in males [27]. In light of our results, we argue that changes in these acoustic features may be capturing deficits in social functioning, explaining their relevance for conversion prediction.

To compare whether the clinical variables (sub and total scores) are useful for predicting psychosis, we used the same classifiers in the three analyzed categories (All, corrected by gender, and only Males). It can be observed that results are at chance level with AUC values even below that 50%. As an additional experiment, we merged our proposed features with clinical variables to see if they can complement each other. Results show inferior performance than using only the proposed features. In fact, the best performance was obtained for only males (AUC = 0.81, accuracy = 0.91 with logistic regression). This means that both types of features are not complementary for prediction of psychosis.

Even though a cross-validated classification model could not be done for females (only one female that transitioned to psychosis), we factorized the features into three dimensions using singular value decomposition (SVD). Figure 1 shows the results of this projection into the components one and two. It can be observed that the female that converted (dot in orange) can be linearly separated. However, we observe that there is not a marked difference with respect to the other CHR- subjects. This could be due to the fact that the conversion for this particular subject happened 16 months from the baseline.

Table 2: Classification results CHR+ vs. CHR-

Type Features	Task	Accuracy	AUC
Proposed Features	**All** (chance: Acc.=0.84/AUC=0.50)	0.84[1] 0.84[2]	0.50[1] 0.50[2]
	All corrected by gender (chance: Acc.=0.84/AUC=0.50)	0.81[1] **0.91[2]**	**0.72[1]** **0.75[2]**
	Males (chance: Acc.=0.81/AUC=0.50)	**0.95[1]** **0.86[2]**	**0.99[1]** **0.87[2]**
Clinical variables	**All** (chance: Acc.=0.84/AUC=0.50)	0.84[1] 0.78[2]	0.43[1] 0.55[2]
	All corrected by gender (chance: Acc.=0.84/AUC=0.50)	0.84[1] 0.81[2]	0.43[1] 0.48[2]
	Males (chance: Acc.=0.81/AUC=0.50)	0.81[1] 0.76[2]	0.40[1] 0.46[2]

[1]Logistic Regression, [2]RBF-SVM

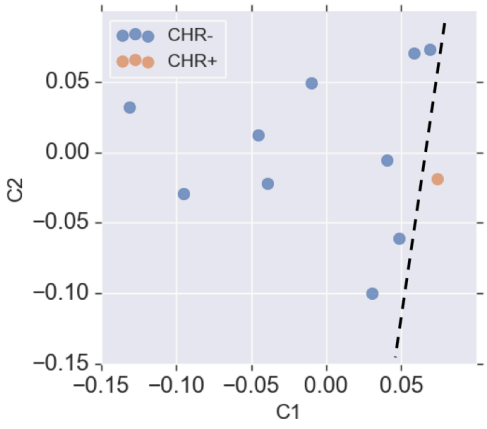

Figure 1 Projection of the proposed features of only females into SVD components.

Finally, the acoustic features of speech identified in this study are potentially early linguistic markers of psychosis risk and, thus could be used for early identification and

longitudinal tracking of individuals that are experiencing psychotic-like episodes. With the advent of smart phones and video messaging, language is easy to capture and repeatable in a way that many neuropsychiatric tests are not. Natural language processing could potentially provide a way to objectively analyze disease progression in "real time" unobtrusively and inexpensively.

V. Conclusions

This study is one of the first of its kind to use automated analysis of acoustic measures of speech to predict transition to psychosis in a CHR cohort. Our results showed that acoustic and prosody features can help to automate the detection of negative symptoms of the SIPS/SOPS with high reliability. In addition, these features can capture subtle changes and include the possibility to analyze longitudinal changes, that can perhaps be missed by humans, to predict psychosis events more accurately.

Limitations of this analysis are the small number of samples, and the use of simple summary metrics (i.e. standard deviation) for quantification of trajectories. Future work will include more samples, which are currently being collected. In addition, a deeper analysis of the trajectories will be done to find clinical subtypes within CHR due to its heterogeneous manifestation.

References

[1] L. A. Schrimpf, A. Aggarwal, and J. Lauriello, "Psychosis," *Contin. Lifelong Learn. Neurol.*, vol. 24, no. 3, BEHAVIORAL NEUROLOGY AND PSYCHIATRY, pp. 845–860, 2018.

[2] H. Killaspy *et al.*, "The impact of psychosis on social inclusion and associated factors," *Int. J. Soc. Psychiatry*, 2014.

[3] J. McGrath, S. Saha, D. Chant, and J. Welham, "Schizophrenia: a concise overview of incidence, prevalence, and mortality," *Epidemiol. Rev.*, vol. 30, no. 1, pp. 67–76, 2008.

[4] C. Hollis and L. Palaniyappan, "Schizophrenia and psychosis," in *Rutter's Child and Adolescent Psychiatry*, Wiley-Blackwell, 2015, pp. 774–792.

[5] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorder Fouth Edition (DSM-IV)*. 2012.

[6] F. Ringeval *et al.*, "AVEC 2017 - Real-life depression, and affect recognition workshop and challenge," in *AVEC 2017 - Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, co-located with MM 2017*, 2017.

[7] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 2014.

[8] C. M. Corcoran, F. Carrillo, D. Fern, G. Bedi, C. Klim, and D. C. Javitt, "Prediction of psychosis across protocols and risk cohorts using automated language analysis," no. February, pp. 67–75, 2018.

[9] G. Bedi *et al.*, "Automated analysis of free speech predicts psychosis onset in high-risk youths," *Nat. Publ. Gr.*, no. June, 2015.

[10] A. S. Cohen, K. R. Mitchell, and B. Elvevåg, "What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments," *Schizophrenia Research*. 2014.

[11] F. Bernardini *et al.*, "Associations of acoustically measured tongue/jaw movements and portion of time speaking with negative symptom severity in patients with schizophrenia in Italy and the United States," *Psychiatry Res.*, vol. 239, pp. 253–258, 2016.

[12] M. A. Covington *et al.*, "Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders," *Schizophr. Res.*, 2012.

[13] E. Caletti *et al.*, "Prosody abilities in a large sample of affective and non-affective first episode psychosis patients," *Compr. Psychiatry*, vol. 86, pp. 31–38, 2018.

[14] T. J. Miller *et al.*, "SYMPTOM ASSESSMENT IN SCHIZOPHRENIC," vol. 70, no. 4, pp. 273–287.

[15] R. Corretge, "Praat vocal toolkit: overview.," 2012. [Online]. Available: http://www.praatvocaltoolkit.com/2012/11/cut-pauses.html. [Accessed: 05-Jan-2017].

[16] J. Zhang, Z. Pan, C. Gui, J. Zhu, and D. Cui, "Clinical investigation of speech signal features among patients with schizophrenia," *Shanghai Arch. Psychiatry*, 2016.

[17] T. Giannakopoulos, "PyAudioAnalysis: An open-source python library for audio signal analysis," *PLoS One*, 2015.

[18] M. Goudbeek, J. P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 1575–1578.

[19] S. Yildirim, M. Bulut, and C. Lee, "An acoustic study of emotions expressed in speech.," *Proc. InterSpeech*, pp. 2193–2196, 2004.

[20] A. S. Cohen, T. L. Renshaw, K. R. Mitchell, and Y. Kim, "A psychometric investigation of 'macroscopic' speech measures for clinical and psychological science," *Behav. Res. Methods*, 2016.

[21] N. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically.," *Behav. Res. Methods*, vol. 41, no. 2, pp. 385–90, 2009.

[22] U. Heitz *et al.*, "Gender differences in first self-perceived signs and symptoms in patients with an at-risk mental state and first-episode psychosis," *Early Interv. Psychiatry*, 2017.

[23] L. Rietschel *et al.*, "Clinical high risk for psychosis: gender differences in symptoms and social functioning," *Early Interv. Psychiatry*, vol. 11, no. 4, pp. 306–313, 2017.

[24] C. M. Corcoran *et al.*, "The relationship of social function to depressive and negative symptoms in individuals at clinical high risk for psychosis," *Psychol. Med.*, 2011.

[25] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, "On the relative importance of vocal source, system, and prosody in human depression," in *2013 IEEE International Conference on Body Sensor Networks, BSN 2013*, 2013.

[26] S. Scherer, G. M. Lucas, J. Gratch, A. Rizzo, and L. P. Morency, "Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews," *IEEE Trans. Affect. Comput.*, 2016.

[27] D. J. Walder *et al.*, "Sexual dimorphisms and prediction of conversion in the NAPLS psychosis prodrome," *Schizophr. Res.*, vol. 144, no. 1–3, pp. 43–50, 2013.