Multimodal Dataset of Social Skills Training in Natural Conversational Setting

Takeshi Saga
NAIST
Ikoma, Nara, Japan
saga.takeshi.sn0@is.naist.jp

Hiroki Tanaka
NAIST
Ikoma, Nara, Japan
hiroki-tan@is.naist.jp

Hidemi Iwasaka
Nara Medical University
Nara, Japan

Yasuhiro Matsuda
Nara Medical University
Nara, Japan

Tsubasa Morimoto
Nara Medical University
Nara, Japan

Mitsuhiro Uratani
Nara Medical University
Nara, Japan

Kosuke Okazaki
Nara Medical University
Nara, Japan

Yuichiro Fujimoto
NAIST
Ikoma, Nara, Japan
yfujimoto@is.naist.jp

Satoshi Nakamura
NAIST
Ikoma, Nara, Japan
s-nakamura@is.naist.jp

ABSTRACT

Social Skills Training (SST) is commonly used in psychiatric rehabilitation programs to improve social skills. It is especially effective for people who have social difficulties related to mental illnesses or developmental difficulties. Previous studies revealed several communication characteristics in Schizophrenia and Autism Spectrum Disorder. However, a few pieces of research have been conducted in natural conversational environments with computational features since automatic capture and analysis are difficult in natural settings. Even if the natural data collection is difficult, the data clearly have much better potential to identify the real communication characteristics of people with mental difficulties and the interaction differences between participants and trainers. Therefore, we collected a one-on-one SST multimodal dataset to investigate and automatically capture natural characteristics expressed by people who suffer from such mental difficulties as Schizophrenia or Autism Spectrum Disorder. To validate the potential of the dataset, using partially annotated data, we trained a classifier for Schizophrenia and healthy control with audio-visual features. We achieved over 85% accuracy, precision, recall, and f1-score in the classification task using only natural interaction data, instead of data captured in the specific tasks designed for clinical assessments.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; • **Applied computing** → *Health care information systems.*

ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8471-1/21/10…$15.00
https://doi.org/10.1145/3461615.3485425

KEYWORDS

Social Skills Training, Schizophrenia, Autism Spectrum Disorder, Audio-visual features

ACM Reference Format:
Takeshi Saga, Hiroki Tanaka, Hidemi Iwasaka, Yasuhiro Matsuda, Tsubasa Morimoto, Mitsuhiro Uratani, Kosuke Okazaki, Yuichiro Fujimoto, and Satoshi Nakamura. 2021. Multimodal Dataset of Social Skills Training in Natural Conversational Setting. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3461615.3485425

1 INTRODUCTION

Social Skills Training (SST), which has been used in clinical populations for over 40 years, was developed based on conditioned reflex therapy, psychotherapy by reciprocal inhibition, and social learning theory [5, 20, 29]. SST improves the social skills that are necessary for living socially at school or in company. SST can be done with one trainer and one participant, or one trainer and several participants. In this paper, the target is one trainer and one participant.

Figure 1 shows SST's basic training flow. The trainer and the participant decide objective skills and the goal of SST. Then, the trainer demonstrates a good example of the skill/goal by acting out the situation him/herself. After that, the participant imitates the trainer's example. Then, positive or negative feedback is given by the trainer. Based on the feedback, the participant repeats his/her performance and tries to improve it if it is recommended. Homework is also assigned to participants to apply the trained skills for daily communication; this process is called generalization.

Although SST is a well-known rehabilitation program that is frequently used in medical fields, the amount of research on the computational analyses of SST interactions is limited [11, 25, 26]. However, we believe that computational analyses are required to find objective and observable features for a more precise evaluation of participants' social skills because such evaluations remain heavily subjective. Actually, these computational approaches can be applied

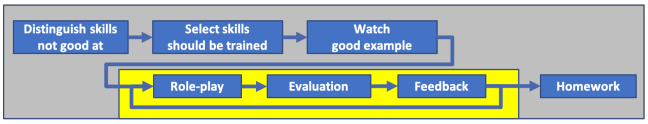

Figure 1: SST Flowchart

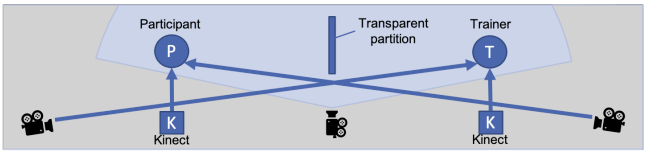

Figure 2: Recording device map

not only for participants' evaluations but those for trainers, too. Currently, there is no concrete evaluation method exists to develop managing skills for trainers. Therefore, they must be evaluated by other trainers. However, the evaluation axes greatly vary from trainer to trainer even if the evaluations are done by experienced senior trainers. For better and more concrete evaluations, we have been working on this topic in a series of our researches [19, 25, 26].

In this paper, as a next step, we collected human-to-human SST data with subjects of Autism Spectrum Disorder (ASD), Schizophrenia (SZ), and healthy control. Moreover, we clarified the significant features in SZ and ASD, which have never been captured or analyzed in natural interactive SST settings. We recorded the interactions with multimodal capture devices for precise analyses. Note that we only executed the yellow part in Figure 1 to control the scenario for our research purposes.

2 RELATED WORKS

Hoque et al. collected job interview data to develop an automated virtual agent called MACH [11], which asks a variety of questions with such interactive responses as nodding or mirroring smiles followed by feedback. Throughout their research, they showed the effectiveness of their interactive virtual agent and its feedback in job interviews. Their team applied the method into the SST context for elderly people and young people in their later works [1, 2].

Voleti et al. collected a dataset of three role-playing scenes [28]. They collected 87 clinical subjects (44 bipolar-i disorder, 43 Schizophrenia or schizoaffective disorders) and 22 healthy controls that participated in the SSPA tasks described by Patterson et al. [18]. They achieved 0.960 ROC scores just using text features. However, non-verbal skills should be considered for fair evaluations.

Compared to previous researches by Hoque and Voleti and our team, we targeted the identification of characteristic multimodal features across SZ or ASD. Moreover, we collected the multimodal data of the subjects and the trainers (conversational opponents), including role-plays and feedback phases, to identify the significant features of better trainers, which is currently unclear. Since this study focuses on multimodal behaviors and interactions of human-human SST, we fixed most SST settings such as feedback strategies across all sessions. Therefore, we won't discuss method differences between trainers or sessions.

3 DATASET

3.1 Setting

All of the data collection processes were approved by ethical committees in Nara Medical University and Nara Institute of Science and Technology. At the beginning of the recording, we explained the procedure to the participants and got informed consent.

Figure 2 shows the recording device map where *P* stands for the participant, *T* stands for the trainer. We used three video cameras and two Azure Kinects (*K* in Figure 2) to capture subtle social signal of the participants and the trainers. Due to the Covid-19 pandemic, we set transparent partitions between the participants and the trainers to prevent infections. We placed two cameras diagonally to avoid light reflection from the partitions to capture facial expressions. Another camera was placed in the center to capture the entire picture. We used Kinects to capture body movements with a pretrained pose estimator based on depth images [24].

To investigate the interaction differences between different types of mental or developmental difficulty, we recruited ASD and SZ sufferers and healthy control subjects. Although equalizing the environments across all the recordings is preferred, our recordings were done at three different places [1] to minimize participants' mental workload. However, we paid much effort to equalize the other environments as much as possible. Among many social skills, we targeted four basic ones: listening, expressing positive or negative feelings, refusing offers, and asking for favors. They are known as the most fundamental skills defined by Bellack [6]. We also used the same workflow and devices across every data collection.

We collected 49 pieces of the adult subjects' data: 19 control (10 male and 9 female. The average age was 28.42. The standard deviation of the age was 3.95), 15 ASD (10 male and 6 female. The average age was 26.50. The standard deviation of the age was 5.67), and 15 SZ (7 male and 8 female. The average age was 32.07. The standard deviation of the age was 8.82). We also collected 33 pieces of children's data for a better investigation of ASD. However, in this paper, we mainly describe the adult subjects. All of the videos were synchronized by a human annotator based on the handclap sign at the beginning and the end of each role-play. All the utterances were transcribed by a human annotator. In addition, we conducted another recording on eye-gaze differences with the same participants on the same day for future research, omitted here due to space limitations. Note that since every subject was a Japanese native speaker, the target language was Japanese.

3.2 Clinical assessments

We gathered several evaluation metrics to clarify the characteristics of the subjects. For this dataset, we collected the Face Emotion Identification Test (FEIT) [14], the Kikuchi's Scale of Social Skills: 18 items (KiSS-18) [15], Singelis scales for Independent-Interdependent self-construal (Singelis) [22], the second edition of the Social Responsiveness Scale (SRS-2) [7], the second edition of the Autism Diagnostic Observation Schedule (ADOS-2) [16], the Japanese version of the Brief Assessment of Cognition in Schizophrenia (BACS-J)

[1]Nara Medical University Hospital, Heart-land Shigisan Clinic, and Nara Institute of Science and Technology

Table 1: Evaluation metrics

Group	Control-adult	ASD-adult	Schizophrenia
FEIT	✓	✓	✓
Kiss-18	✓	✓	✓
Singelis	✓	✓	✓
SRS-2	✓	✓	
ADOS-2		✓	
BACS-J		✓	✓
PANSS			✓

[13], and the Positive and Negative Syndrome Scale (PANSS) [12], depending on each symptomatic group. Table 1 shows the correspondence between the metrics and the groups.

FEIT, which assesses the emotional perception of facial emotions, includes the facial images in greyscale of 19 different people with one of six emotions: happiness, sadness, anger, surprise, fear, and shame. We included it since ASD struggle with emotion recognition from facial images [9]. Kiss-18, which measures social skills levels, is composed of 18 questions based on six social skill categories defined by Goldstein [10]. This metric comprehensively measures social skills. Singelis is constructed of 30 questions on a 7-point rating scale. It was developed to measure how people view themselves to others. SRS-2 is an evaluation metric for the severity of social impairment that is composed of 65 questions. Although SRS-2 was originally designed to assess potential ASD sufferers, it can also differentiate among a variety of mental difficulties. Since its effectiveness has been investigated with the challenged and the healthy, it is suitable for evaluating healthy people as well [7]. ADOS-2 is a semi-structured assessment that includes several play-based activities for collecting information related to communication, social interactions, and restricted and repetitive behaviors associated with ASD. Since it doesn't depend on any language levels, it can be uniformly applied to kindergarten children and adults. BACS-J is a Japanese version of the Brief Assessment of Cognition in Schizophrenia (BACS) that assesses the aspects of cognition found to be most impaired and most strongly correlated with outcomes in SZ patients. Since not only SZ but also ASD include cognitive impairments, we also applied BACS-J to ASD subjects. PANSS is an instrument for typological and dimensional assessment for SZ that is composed of standardized 30 items. It provides a balanced representation of positive and negative symptoms. With these metrics, we plan to identify the characteristic differences among the clinical groups. Our annotator is currently compiling these metrics.

Since we tried a classification between SZ and control subjects in section 4, we investigated the assessment score differences between them. With the Mann-Whitney U rank test, we confirmed that there weren't significant differences ($p>0.05$) for Kiss-18, Singelis, and FEIT with p-values 0.94, 0.16, and 0.39 respectively.

3.3 Subjective evaluations

Furthermore, we subjectively evaluated the social skills of participants and trainers by other experienced trainers by watching the videos afterward. We evaluated the participant acts with the following seven components: eye contact, body direction and distance, facial expression, voice variation, clarity, fluency, social appropriateness for each task. In contrast, we evaluated the trainer acts with the following four components: appropriateness of positive feedbacks, suggestions for improvement, appropriate non-verbal communication as a good example, appropriate verbal communication as a good example. Every component was evaluated with a 5-point Likert scale. Each video was evaluated by two evaluators.

To validate the evaluation scores' reliability, we calculated Cohen's quadratic kappa scores. The reliability was confirmed with the kappa score of 0.84 across entire subjective evaluations. There were no significant differences in kappa scores between different SST tasks. On the other hand, healthy controls were comparably lower than SZ, and SZ was lower than ASD. We believe this indicated that the variability of social skills levels of healthy controls was more diverse than other groups. Similarly, the kappa score for the evaluations on participants was lower than the one on trainers.

Table 2: Audio features

Feature name	Description
F0, F1, F2, F3 Mean	Mean frequency of F0, F1, F2, F3
F0, F1, F2, F3 SD	Standard deviation of F0, F1, F2, F3
F0 Min, Max	Minimum and Maximum F0 frequency
F0 range	Range (Max - Min) of F0
F1, F2, F3 BW	Average bandwidth of F1, F2, F3
F2/F1, F3/F1 Mean	Mean ratio of F2-F1 and F3-F1
F2/F1, F3/F1 SD	Standard deviation of F2/F1 and F3/F1
Int range	Range (Max - Min) of intensities
Int SD	Standard deviation of vocal intensity

4 CLASSIFICATION OF SCHIZOPHRENIA AND HEALTHY CONTROL

We validated the effectiveness of this dataset by training a classifier using multimodal features. As an experimental study, we set the adult control subjects and the adult SZ subjects as the classification targets. Unfortunately, we couldn't include the ASD subjects in this validation since we haven't finished the annotation yet.

4.1 Method

We used the random forest as the machine learning model. For its input, we used audio-visual features calculated by existing software and preprocessed the data with the following procedure. First, audio data were separated from the video data. Then, we separated the data into several segments according to SST phases, such as role-plays or feedbacks. For the audio data, next, we eliminated unvoiced and overlapped voiced segments between the trainer and the participant. After that, we concatenated non-overlapped voiced segments. Since the several files were too short for accurate calculation, we eliminated audio files shorter than five seconds. Note that we didn't use body joint position since we have not found a way to synchronize the data at the boundary point of SST phases yet.

We used Praat for the audio feature calculation and OpenFace for the visual feature calculation [3, 4, 27]. Tables 2 and 3 show the input features, where F0 indicates the fundamental frequency of voiced segments, F1 to F3 indicate the first to the third formant of

Table 3: Visual features

Feature name	Description
Pose_[Rx,Ry,Rz]_std	Standard deviation of head rotation around (X:pitch,Y:yaw,Z:roll) axis
Smile_ratio	Ratio of Smiling frames
AU01	inner-brow raiser
AU02	outer-brow raiser
AU04	brow lowerer
AU05	upper-lid raiser
AU06	cheek raiser
AU07	lid tightener
AU09	nose wrinkler
AU10	upper-lip raiser
AU12	lip-corner puller
AU15	lip-corner depressor
AU17	chin raiser
AU20	lip stretcher
AU23	lip tightener
AU25	lips parter
AU26	jaw dropper
AU28	lip sucker
AU45	blinker

voiced segments, AUXX indicates facial Action Unit (AU) which is a component of each facial expression. Since the background noise of the locations was different, perhaps the audio intensity changed depending on the locations. Therefore, we didn't include the average, minimum, and maximum values of the intensity to avoid the location differences, although the range and standard deviation were included because they were potentially less affected.

Before being input into the model, all features were standardized to have the mean of zero and the standard deviation of one. To investigate which modality is dominant for this classification, we set three different feature-set conditions: audio-only (A), visual-only (V), audio-visual (A+V). We trained the model ten times with random parameters and averaged the results to get a final result.

4.2 Results and discussion

Table 4 shows the classification results. The highest F1-score was 0.887. In terms of modality differences, the audio-only features showed the best performance across all the metrics. In contrast, the visual-only features showed significantly lower performance than the audio-only. The audio-visual features showed an average performance. These classification results indicated the importance of audio features to differentiate two groups. Actually, prior research reported vocal atypicalities in SZ [17]. Our classification results with different modality-sets supported them.

Furthermore, we investigated the feature importances for the classification with random forest. Table 5 shows the top-3 feature importances for the best model (audio-only). *F3 BW* shows the highest importance value with over double the value of the others. It is acknowledged that wide formant bandwidth induces a significant reduction of the vowel identification rate [8, 21, 23]. In fact, we

Table 4: Classification results

Feature	Accuracy	Precision	Recall	F1-score
A	0.868	0.902	0.872	0.887
V	0.382	0.476	0.488	0.422
A+V	0.804	0.823	0.851	0.836

Table 5: Top-3 important features (audio-only)

Feature name	F3 BW	F2/F1 Mean	F1 BW
Importance	0.231	0.089	0.087

confirmed that the bandwidth of the SZ group was mostly wider than the one of the control group by comparing each feature value. Therefore, our result indicates the SZ group participants tend to have less vowel intelligibility than the control group, which can be a part of communication problems.

5 CONCLUSION AND FUTURE WORK

We explained our SST data collection including subjects with mental or developmental difficulty and classification with audio-visual features. Our classification demonstrated the usefulness of our natural conversational SST dataset. Although researchers have tried to classify SZ in limited experimental situations, our classification with this natural dataset achieved reasonably high performance.

Toward an automated SST system, several challenges remain. First, an analysis of the trainers' behavior is necessary to achieve effective feedbacks. Such an idea is meaningful not only to develop better automated SST trainers but objectively measure trainers' skills, which is still difficult. Second, further investigations are needed of the characteristic features of SZ. For years, trainers clearly recognize the existence of praecox feelings when they talk with SZ. However, trainers couldn't explain the feelings in detail because it was ambiguous. We believe that our computational analyses will provide answers to that question in the future.

As a limitation of our research, we couldn't equalize relationships between the trainers and the subjects. Since we couldn't find any other way to recruit participants with difficulties, we recruited them from the hospitals at which the trainers were working. Therefore, the trainers and the disordered subjects had relationships before the recording, whereas the trainers and the control subjects didn't have since they were located through a subject recruiting service. Such differences in locations and relationships are probably a limitation of our current research. In addition, we recruited four male and one female trainers for this research. Note that the results might be affected by the imbalanced gender ratio of trainers.

ACKNOWLEDGMENTS

Funding was provided by the Core Research for the Evolutional Science and Technology (Grant No. JPMJCR19A5) and the Japan Society for the Promotion of Science (Grant Nos. JP17H06101 and JP18K11437). We appreciate grateful help to other assistants in the data collection, including laboratory co-workers and experienced SST trainers.

REFERENCES

[1] Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K. Schubert, and Mohammad Ehsan Hoque. 2020. A Virtual Conversational Agent for Teens with Autism Spectrum Disorder: Experimental Results and Design Lessons. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20)*. Association for Computing Machinery, New York, NY, USA, Article 2, 8 pages. https://doi.org/10.1145/3383652.3423900

[2] Mohammad Rafayet Ali, Kimberly Van Orden, Kimberly Parkhurst, Shuyang Liu, Viet-Duy Nguyen, Paul Duberstein, and M. Ehsan Hoque. 2018. Aging and Engaging: A Social Conversational Skills Training Program for Older Adults. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, New York, NY, USA, 55–66. https://doi.org/10.1145/3172944.3172958

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 59–66.

[4] T. Baltrušaitis, M. Mahmoud, and P. Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–6.

[5] A. Bandura. 1969. *Principles of behavior modification*. Holt, Rinehart and Winston.

[6] Alan S. Bellack, Kim T. Mueser, Susan Gingerich, and Julie Agresta. 2004. *Social Skills Training for Schizophrenia: A Step-by-Step Guide* (2 ed.). Guilford Press.

[7] John N. Constantino and Christian P. Gruber. 2012. *Social Responsiveness Scale, SRS-2* (2 ed.). Western Psychological Services.

[8] J. R. Dubno and M. F. Dorman. 1987. Effects of spectral flattening on vowel identification. *J Acoust Soc Am* 82, 5 (Nov 1987), 1503–1511.

[9] O. Golan, S. Baron-Cohen, J. J. Hill, and Y. Golan. 2006. The "reading the mind in films" task: complex emotion recognition in adults with and without autism spectrum conditions. *Soc Neurosci* 1, 2 (2006), 111–123.

[10] I.L. Goldstein. 1986. *Training in Organizations: Needs Assessment, Development and Evaluation*. Brooks/Cole publishing company.

[11] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: My Automated Conversation Coach. In *Proceedings of UbiComp '13*. Association for Computing Machinery, New York, NY, USA, 697–706. https://doi.org/10.1145/2493432.2493502

[12] S. R. Kay, A. Fiszbein, and L. A. Opler. 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull* 13, 2 (1987), 261–276.

[13] R. S. Keefe, T. E. Goldberg, P. D. Harvey, J. M. Gold, M. P. Poe, and L. Coughenour. 2004. The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophr Res* 68, 2-3 (Jun 2004), 283–297.

[14] S. L. Kerr and J. M. Neale. 1993. Emotion perception in schizophrenia: specific deficit or further evidence of generalized poor performance? *J Abnorm Psychol* 102, 2 (May 1993), 312–318.

[15] Akio Kikuchi. 1988. The development of a social skills scale. 38 (1988), 67–68. In Japanese.

[16] Catherine Lord and Michael Rutter. 2012. *Autism Diagnostic Observation Schedule, Second Edition*. WPS.

[17] A. Parola, A. Simonsen, V. Bliksted, and R. Fusaroli. 2020. Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. *Schizophr Res* 216 (02 2020), 24–40.

[18] T. L. Patterson, S. Moscona, C. L. McKibbin, K. Davidson, and D. V. Jeste. 2001. Social skills performance assessment among older patients with Schizophrenia. *Schizophrenia Research* 48, 2–3 (3 2001), 351–360.

[19] Takeshi Saga, Hiroki Tanaka, Hidemi Iwasaka, and Satoshi Nakamura. 2020. Objective Prediction of Social Skills Level for Automated Social Skills Training Using Audio and Text Information. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 467–471. https://doi.org/10.1145/3395035.3425221

[20] A. Salter. 1949. *Conditioned reflex therapy*. Creative Age Press.

[21] S Sekimoto. 1982. Effects of formant peak emphasis on vowel intelligibility in frequency compressed. *Annual Bulletin of logopedics and phoniatrics* 16 (1982).

[22] Theodore M. Singelis. 1994. The Measurement of Independent and Interdependent Self-Construals. *Personality and Social Psychology Bulletin* 20, 5 (1994), 580–591. https://doi.org/10.1177/0146167294205014

[23] Quentin Summerfield, John Foster, Richard Tyler, and Peter J. Bailey. 1985. Influences of formant bandwidth and auditory frequency selectivity on identification of place of articulation in stop consonants. *Speech Communication* 4, 1 (1985), 213–229. https://doi.org/10.1016/0167-6393(85)90048-2

[24] T. Sych, C. Casey, and P. Meadows. [n.d.]. Azure Kinect DK Documentation. https://docs.microsoft.com/en-us/azure/kinect-dk/ Last accessed: August 2021.

[25] Hiroki Tanaka, Hidemi Iwasaka, Hideki Negoro, and Satoshi Nakamura. 2020. Analysis of conversational listening skills toward agent-based social skills training. *Journal on Multimodal User Interfaces* 14, 1 (01 Mar 2020), 73–82. https://doi.org/10.1007/s12193-019-00313-y

[26] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2017. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLOS ONE* 12, 8 (08 2017), 1–15. https://doi.org/10.1371/journal.pone.0182151

[27] Vincent van Heuven. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 9/10 (2001), 341–345.

[28] Rohit Voleti, Stephanie Woolridge, Julie Liss, Melissa Milanovic, Christopher Bowie, and Visar Berisha. 2019. Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder. In *Proceedings of Interspeech 2019*. International Speech Communication Association, 1433–1437. https://doi.org/10.21437/Interspeech.2019-2960

[29] J. Wolpe. 1958. *Psychotherapy by reciprocal inhibition*. Stanford University Press.