Predicting Self-Reported Affect from Speech Acoustics and Language

Chelsea Chandler[1,2], Peter W. Foltz[2,3], Jian Cheng[4], Alex S. Cohen[5], Terje B. Holmlund[6], Brita Elvevåg[6,7]

[1] Department of Computer Science, University of Colorado Boulder, CO, U.S.A.
[2] Institute of Cognitive Science, University of Colorado Boulder, CO, U.S.A.
[3] Pearson, CO, U.S.A. [4] Analytic Measures Inc., Palo Alto, CA, U.S.A.
[5]Department of Psychology, Louisiana State University, LA, U.S.A.
[6] Department of Clinical Medicine, UiT The Arctic University of Norway, Norway
[7] Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway
{chelsea.chandler, peter.foltz}@colorado.edu, brita@elvevaag.net

Abstract

Communication via speech offers a window into a person's mental and cognitive states. Both the manner in which a person speaks (acoustics) and the words spoken (language) may be used to assay current mental and cognitive function. In this study, we predicted self-reported emotion from the acoustics and language of the seemingly affectless task of verbally recalling a short story. Story recalls and self-reports of affect were collected over multiple days via a mobile application in a population of 21 psychiatric patients and 79 presumed healthy participants, resulting in 137 and 430 total sessions for each group respectively. We have previously shown that analyzing just one modality of data produces moderate correlations with self-reported affect ($0.33 < r < 0.40$ for speech and $0.07 < r < 0.28$ for language). The goal of this study was to improve on unimodal analyses by extracting acoustic and language features from story recalls and combining them to predict a person's self-reported affect. This combination of modalities resulted in an improvement over just one modality alone ($0.38 < r < 0.48$). We show that a multimodal analytic approach predicted self-reported emotional states in clinical and non-clinical participants better than a unimodal approach.

Keywords: positive affect, negative affect, speech, natural language processing, machine learning

1. Introduction

Psychosis can disrupt language production in a number of different ways, including acoustics (transmitted sound), lexemes (word choice), syntax (sentence structure), coherence (logical flow), and semantics (meaning; see Holmlund et al., 2020b for a review). Therefore, the evaluation of language production is an important component during clinical interviews and in many standard psychosis rating scales. Such formal examinations would hugely benefit from more objective and rigorously defined analyses where automation can improve speed, reliability and consistency of judgments. In this study, we first sought to answer whether machine learning techniques could infer what aspects of speech acoustics and what words and patterns of words in language could be used to predict positive and negative affect, and second whether the combination of these two modalities would improve on unimodal predictions.

Clinically valuable characteristics, notably negative affect, exist even just in the sound of patients' voices (Cohen et al., 2016a, 2016b). An example of such a characteristic is a lack of vocal modulation across changing clinical state. The clinical value in such a feature is that it can be a potential indicator of a worsening clinical state. Therefore, recordings of patients' speech can be a critical component in the monitoring of patients with serious mental illness as these would allow clinicians to more accurately track dynamic signals over time in conjunction with other state-related variables in formal analyses (Cohen et al., 2019).

Linguistic variables of patient speech have been shown to contain power in predicting variables of interest in numerous clinical settings. For instance, Elvevåg et al. (2007) showed that levels of incoherence in language can be used in differentiating diagnostic groups and detecting severity in schizophrenia. Recent work in the field of artificial intelligence, and more specifically natural language processing (NLP), has shown that various cognitive variables can be predicted from language alone (for a review, see Voleti et al., 2019).

Clinicians must integrate information from various behavioral, self-report, and historical sources during the assessment process. The language component of such assessments is just one of a multitude of modalities evaluated. Numerous aspects of serious mental illness can be conveyed in subtleties in a patient's vocal self-presentation including emotions as expressed in both the sound of voice and the types of words used (Holmlund et al., 2020b), and clinicians may have difficulty noticing and remembering these distinctions. Mental disorders require longitudinal monitoring over several years which is extremely challenging cognitively for clinicians given the different clinical baselines of patients with serious mental illness. Furthermore, it is a difficult task for a human to evaluate the degree to which single modalities of behavior contribute to an individual's overall mental and cognitive health. Thus, there is a pressing need for automated analytic methods to track and assess mental state in a scalable manner.

As part of a larger study, data were collected through a mobile phone application, the *delta* Mental State Examination (*d*MSE), which administered assessments and collected speech, touch, and self-report data from users in order to track changes in mental state over time (Holmlund et al., 2019; Cohen et al., 2019). Of the 12 tasks administered to the users of the application, we chose a story recall task to answer the question of whether it is possible to predict self-reported measures of positive and negative affect using both speech acoustics and language features.

2. Related Work

It has been shown that emotional state can be automatically measured through a person's speech, both in and out of the laboratory. For instance, the studies of Grimm et al. (2007), Asgari et al. (2014), and more recently, Zhaocheng and

Epps (2018), showed that spontaneous emotion could be accurately predicted from speech. However, all studies took place in a controlled laboratory setting. Our recent work has shown that such analyses are also viable in less controlled settings when tasks are administered remotely via a mobile application (Cheng at al., 2018). In the study of Cheng et al. (2018), 10 speech-based tasks were used to predict self-reported positive and negative affect using only acoustic features. In the current study, we extended this to two modalities: speech acoustics and language, while focusing our analyses only on one single task that does not explicitly elicit emotion (story recall).

One way to measure emotional state is by viewing it as a classification task. In such a task, the goal is to predict speech as belonging to one of the basic categories of emotion (e.g., happiness, anger, fear, etc.). This approach can be problematic since it is hard to get reliable categorization of emotion across evaluators (Mower et al., 2009). While some studies of emotional prediction have attempted to mitigate and overcome this drawback (Steidl et al., 2005), most focus on the prediction of the extent to which certain categories of emotion are present via a continuous representation (Cowie et al., 2012). Thus, the present study employed a regression model to predict emotional state.

Similar to the aforementioned studies, much of the analyses and modeling of behavioral and psychiatric data to date has been in a unimodal manner. For instance, the Interspeech 2018 Computational Paralinguistics Challenge aimed to increase the sensitivity to the non-language information that is conveyed in acoustic properties of speech. Specifically, the self-assessed affect sub-challenge sought to predict the valence of emotions, with the objective of supporting applications for individuals with affective disorders, and for monitoring interactions between therapists and their patients (Schuller et al., 2018).

By focusing solely on acoustic properties, these studies miss the signal contained in natural language features. While the manner in which language is produced at an acoustic level is decidedly important, classic language features have been shown to serve as a window into a person's mental state. For example, natural language features have been shown to accurately predict performance on a story recall task often given as part of the clinical workup in psychiatric settings (Chandler et al., 2019a; Holmlund et al., 2020a). Natural language features have been studied in a range of clinical applications from detecting language impairments in autism to flagging depression in twitter feeds (Goodkind et al., 2018; Coppersmith et al., 2015).

In each study, patient data was reduced to a set of variables to relate to clinical measures of interest. Whether the modality of choice is acoustics, language, reaction time, precision, etc., it has been shown that psychiatric variables of interest can be accurately predicted from unimodal data.

3. Data Collection

The *d*MSE mobile phone application was created for the acquisition of cognitive and mental health data of various modalities from both a clinical and non-clinical population. Participants remotely completed sessions, each consisting of a series of 12 tasks, over the course of 3 to 6 days. Such tasks were created to be similar in form and structure to those employed by clinicians in standard neuropsychological evaluations. One part of each session prompted participants to answer several questions on their emotional well-being by moving a slider to indicate their current level of positive and negative affect (Cohen et al., 2019; Cowan et al., 2019; Le et al., 2018, 2019; more detail in the next section). For the purpose of this study, slider and story recall results were captured by a smart device running the *d*MSE application.

The story recall task prompted participants to listen to a short story and then retell it immediately in as much detail as possible. Stories were all presented verbally and contained two characters, a setting, an action that caused a problem, and a resolution. The content of the stories were designed to be generally well known topics that were emotionally neutral. On average each story was 72 words (SD = 4.6) and each retell was 61.3 words (SD = 21.2 words) and 41.7 words (SD = 21.0 words) for non-clinical participants and clinical participants, respectively. An example story was as follows:

> "On Monday morning, the woman woke up more tired than usual. When she walked downstairs to make herself a cup of coffee, she found her husband in the kitchen. She was surprised because he usually left an hour before she woke up. Her husband greeted her and reminded her that daylight savings time was over. Realizing the clocks were wrong, she happily ran upstairs and jumped back into bed."

The non-clinical subset of our data was composed of 430 sessions that produced valid data from 79 (presumed healthy) undergraduates enrolled in psychology courses at Louisiana State University, yielding 5.4 sessions per student. The clinical subset of our data was composed of 137 sessions that produced valid data from 21 stable clinical participants with a range of serious mental illnesses (schizophrenia, major depressive disorder and bipolar disorder; for details on the assessment procedure in a slightly extended sample, see Holmlund et al., 2020a), yielding 6.5 sessions per participant. This study was approved by the LSU Institutional Review Board (#3618) and participants provided their informed written consent before participation.

4. Self-Reported Affect

Each session with the *d*MSE application included sliders to assess general affective states. Participants were prompted as to their emotion on various questions (described below) and asked to indicate their responses on a scale of 0-100. The questions were based on the Positive and Negative Affect Schedule (Watson et al., 1988), which is a tool that measures Positive Affect (PA) and Negative Affect (NA). PA is defined as a state of high enthusiasm, activity, and alertness and NA is defined as a state of distress and unpleasurable engagement. Both are used to quantify mood and are known to be relatively independent of one another. The *d*MSE application contains 7 PA sliders that ask the user to report on personal levels of hopefulness, calmness, appreciation, strength, ability to concentrate, happiness, and levels of energy. Similarly, 8 NA sliders ask the user to self-report on personal levels of anxiety, frustration, fear, sadness, stress, anger, pain, and helplessness. The final self-reported PA and NA values per session is the average of the PA and NA slider responses. The PA results ranged

from 0 to 100 with an average of 74.9 (SD = 21.0) for the clinical group and ranged from 10 to 100 with an average of 64.0 (SD = 17.3) for the non-clinical group. The NA results ranged from 0 to 100 with an average of 29.5 (SD = 23.2) for the clinical group and ranged from 0 to 74 with an average of 26.1 (SD = 17.0) for the non-clinical group.

5. Experimental Results

In this study, we generated predictions of positive affect (PA) and negative affect (NA) in a clinical group and a non-clinical group. The predictions were first based on speech features and standard NLP features individually, and then on a combination of these two to answer the question of how well multimodal predictions outperform unimodal predictions.

In each experiment variation, a Support Vector Regression (SVR) model was trained on data from all participants in a group but one, and tested on the set of sessions of each 'held out' participant. The SVR model was chosen as it is well-suited for predicting with many continuous independent variables. The SVR parameters were consistent with prior work: a radial basis function (RBF) kernel, degree = 3, cost = 10, eps = 0.2 (Cheng et al., 2018). The reported results are the average correlation between self-reported PA and NA and the predicted PA and NA over all tested participants. Each model was trained with these same parameters as we were more interested in relative improvements in the overall prediction of PA and NA when new features and modalities were introduced than finding the best overall models.

5.1 Speech-based results

The first experiment was a re-analysis of prior work (see Cheng et al., 2018 for details). Speech features from the openSMILE audio feature extractor (Eyben et al., 2013) were generated from each story recall response. The openSMILE audio feature extractor is a state of the art package that generates low-level features such as energy, loudness, and voice quality as well as processed statistics of such features such as means, extremes, regressions, and percentiles. We used the entire 2013 ComParE feature set which comprised 6,373 distinct speech features per response (Schuller et al., 2013).

Prior work reported results on the same data and model, but used a 10-fold cross-validation training technique where model parameters were learned using 9 of 10 subsets of the data and tested on the 10th subset for evaluating performance. In contrast, we performed a leave-one-out cross-validation technique where the model parameters were learned using data from all but one participant and tested on the set of sessions from the single 'held out' participant. The benefit of this form of cross-validation is that the resulting models more closely resembled how well a fully trained model would perform when applied to new data.

The subsequent analyses were performed on various subsets of the data. All openSMILE features were used in the SVR to predict both PA and NA in the clinical participants as well as in the non-clinical participants. For this part of the analysis, the groups were kept separate. The results of the different variations of the analysis are shown in Table 1. Consistent with prior work (Cheng et al., 2018), we found higher correlations to self-reported affect in the clinical population than the non-clinical population.

	Clinical	Non-clinical
PA	0.40	0.33
NA	0.39	0.36

Table 1: Correlations between self-reported PA/NA and SVR predictions using all openSMILE features.

Many of the features in the openSMILE feature set are highly co-linear, and so receive essentially 0 weighting within the prediction model. While many contribute to the prediction models nearly equally based on small differences in the data, each iteration of the prediction model had a distinct best feature that correlated significantly higher than the rest. For instance, spectral flux and Mel-Frequency Cepstral Coefficients (MFCC) best predicted PA and NA respectively in the clinical group. Similarly, spectral roll-off and MFCC best predicted PA and NA respectively in the non-clinical group.

In addition to models trained and tested on clinical and non-clinical participants separately, a SVR model was trained on data from all clinical participants and tested on data from all non-clinical participants, as well as *vice versa*. The results of this portion of the analysis are detailed in Table 2. When combining data from the two groups of participants, the correlations with PA and NA both significantly decreased. This suggests that the weights of the various speech features used to predict PA and NA have different distributions in the two subsets of participants due to the difference in ranges of self-reported affect.

Training set	Test set	PA	NA
Clinical	Non-clinical	0.16	0.11
Non-clinical	Clinical	0.06	0.08

Table 2: Correlations between self-reported PA/NA and SVR predictions using all openSMLE features when trained on the non-clinical population and tested on clinical, as well as *vice versa*.

Finally, one model with all clinical and all non-clinical data combined was trained and tested in the same leave-one-out manner. The average correlation between the SVR predictions and self-reports of PA was 0.13 and of NA was 0.06. The lower correlations that result from both models that mix clinical and non-clinical data imply that these two populations must be considered independently as their self-reports follow distributions that are distinct from one another.

5.2 Language-based results

Since the aim of this study was to test whether the addition of data from a separate modality would improve the ability to predict emotion, we next repeated the speech-based experiments on language-based features to compute a baseline of the power of language features.

Traditionally, story recall is rated manually by assigning points for key words or thematic units correctly recalled. This process can be automated by extracting various task-specific NLP features (e.g., common tokens between the original story and the recall or the cosine distance between the vector representations of the original story and the recall) from each recall response to measure the similarity

between the two (see Chandler et al., 2019 and Holmlund et al., in press for more details).

The audio of each story recall was transcribed by trained humans. Non-task-specific NLP features were computed and modeled against the self-reported affect variables (so as to focus the analyses only on general language features rather than those features that would indicate successful task completion). The NLP feature set included token count, type (unique words) count, type token ratio, content density, mean coherence, standard deviation of coherence, and counts of particular parts of speech such as verbs, nouns, pronouns. Type token ratio is defined as the ratio of word types to word tokens. Content density is defined as the number of verbs, nouns, adjectives, and adverbs to total tokens, or put simply, the ratio of content words to total words. Coherence is computed by comparing adjacent windows in the text for similarity. For the purpose of this study, the window size was chosen to be n = 4 and the similarity metric used was the cosine distance between vector embeddings of the words in each window. The average and standard deviation of similarities of all adjacent windows in a recall were computed. Table 3 shows the results of the different variations of analyses conducted on language-features (which are identical to the variations in the speech-based experiments).

	Clinical	Non-clinical
PA	0.16	0.07
NA	0.28	0.14

Table 3: Correlations between self-reported PA/NA and SVR predictions using all standard NLP features.

Perhaps unsurprisingly, the NLP features were less useful than the speech features in predicting affect. Had the task analyzed been one that specifically seeks to elicit affect in the words spoken, such as the prompt "how are you feeling today?", we predict that NLP features would be more likely to have a stronger impact on modeling affect.

Finally, a semantic analysis was performed using high-dimensional vector space embeddings of the text. The purpose of the semantic analysis was to measure subtle aspects of language meaning that could be correlated with self-reported affect across different participants. These embeddings operate under the assumption that words that tend to show up in similar contexts are semantically related and thus should be close to each other in a derived vector space. Examples of embedding techniques are simple count based vectorizers (with and without term frequency-inverse document frequency weighting), Latent Semantic Analysis (Landauer and Dumais, 1997), word2vec (Mikolov et al., 2013), and ELMo (Peters et al., 2018). In this experiment, the term frequency-inverse document frequency weighted vectors were the most predictive out of those tested. The term frequency-inverse document frequency weighting accounts for how important a word is to a document based on counts of the word in the entire corpus. Although word2vec and ELMo embeddings are typically regarded as containing more signal in terms of word meaning, they proved unable to predict affect. This is likely due to the fact that the vocabulary of the recall task is limited and thus the increased power of the semantic/syntactic modeling found in these embeddings do not contribute greatly to predicting affective state. The two variations of our two semantic language-based experiments are detailed below.

First, a k-nearest neighbors (KNN) measure was developed to predict PA and NA based on the affect ratings of the closest recalls in the embedding space to a given recall. Once each recall is projected into an embedding space, the k = 6 (chosen based on the overall best performance on the held out data) closest embeddings of other participant recalls were retrieved and the affect for the session in question is predicted to be some function of those 6. The KNN measure provides an indexing of a participant's PA and NA mental state against the mental state of other participants. For example, if the language used in a response is highly similar to other participants, we can predict that their PA and NA scores would be similar.

Second, the recall embeddings were used as input to a SVR model. The same experimental settings were used as in the SVR for speech-based features and the standard NLP-based features. Results of the KNN model and the SVR model are detailed in Table 4. Again, the speech-based features consistently outperform the language-based features.

	Clinical	Non-clinical	Clinical	Non-clinical
	KNN		SVR	
PA	0.11	0.07	0.23	0.18
NA	0.14	0.16	0.31	0.13

Table 4: Correlations between self-reported PA/NA and both KNN predictions and SVR predictions using the recall embedding as input.

5.3 Combined results

Finally, to test our hypothesis that the inclusion of multiple modalities in the modeling of self-reported affect is superior to unimodal modeling, we combined the speech-based features with the language-based features and ran the same SVR experiment variations as above.

Combining two modalities improves predictions of self-reported affect by 10-23%. Even though the recall task is a task that is not designed to specifically elicit emotion, the manner in which participants spoke in terms of acoustics and language still contained critical signals indicative of their positive and negative affect.

All features (openSMILE, standard NLP, and recall embedding neighbors) were used in a single SVR model to predict positive and negative affect. Results of this combined model are detailed in Table 5.

	Clinical	Non-clinical
PA	0.44 (10%)	0.38 (15%)
NA	0.48 (23%)	0.40 (11%)

Table 5: Correlations between self-reported PA/NA and SVR predictions using all features, with their relative improvements over unimodal predictions.

6. Discussion and Conclusion

In this research, self-reported measures of affect (e.g., PA/NA) were taken separately from the story recall task that was used to predict emotion. Indeed, the nature of a story recall task on neutral stories is not directly designed to elicit emotional state. However, subtle aspects of

emotion were still evident in the language. These results indicate that the approach can derive a fairly stable measure of affect through self-reports that can be predicted in separate tasks. Furthermore, the results indicated that some people's affect levels are easier to predict than others and some types of affect may be easier to predict. For example, Cheng et al. (2018) showed that negative affect was easier to predict than positive affect.

Overall, we have shown that the use of multiple modalities of data in prediction models can lead to a significant increase in power over analyses of a single modality. Speech and language features each contribute independent components that help predict affective state. The speech features contributed more strongly to the predictions which could partially be due to the nature of the tasks used.

Traditionally, unimodal data analyses have been conducted on clinically valuable data as the combination of modalities (and thus data types) can be statistically complex. However, the field of clinical medicine and behavioral science is beginning to see a push for more multimodal analyses. Although the collection of multimodal data is standard in many fields (e.g. neuroimaging; Sun et al., 2020), it is just recently becoming common for multiple modalities to be considered in a single computational model.

Overall, the results show a path towards automatic analysis of patient mental state using both audio and linguistic features. This research has shown that affective state can be predicted from a single task with two modes of communication. In automated assessment of psychiatric variables, it is important to consider multiple modalities of behavior, whether that is within language (considering both acoustic and linguistic data), or beyond, using patient actions, response speed, and other similar variables.

The *d*MSE collected data from a variety of other tasks, including picture descriptions, verbal fluency, memory, tapping, and Stroop tasks. Thus, future research will examine the data from all tasks and run equivalent SVR experiments on features extracted from all sessions of each participant. This opens the possibility of analyzing a combination of speech, language, memory accuracy and touch-based speededness tasks, and could give a more detailed and accurate view of the patients' state, more analogous to what a clinician considers when making decisions. However, with the increase in model complexity, we must be careful, especially in the field of medicine, to not lose the notion of transparency and explainability (Chandler et al., 2019b). Thus, while adding additional modalities and features, it is critical to remember that the goal is not just to build an accurate model, but to understand how that model can be used to inform sound clinical decision-making.

7. Acknowledgements

This project was funded by grant 231395 (2014 – 2017) from the Research Council of Norway awarded to Brita Elvevåg. We thank Jared Bernstein and Elizabeth Rosenfeld for their valuable contributions to the project.

8. Bibliographical References

Asgari, M., Kiss, G., van Santen, J., Shafran, I., and Song, X. (2014). Automatic measurement of affective valence and arousal in speech. In Proceedings of ICASSP, pp. 965-969.

Chandler, C., Foltz, P.W., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Cohen, A.S, Holmlund, T.B. and Elvevåg, B. (2019a). Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pp. 137–147. http://dx.doi.org/10.18653/v1/W19-3016

Chandler, C., Foltz, P.W., Elvevåg, B. (2019b). Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness, Schizophrenia Bulletin, Volume 46, Issue 1, January 2020, pp. 11–14, https://doi.org/10.1093/schbul/sbz105

Cheng, J., Bernstein, J., Rosenfeld, E., Foltz, P.W., Cohen, A., Holmlund, T., Elvevåg, B. (2018). Modeling Self-Reported and Observed Affect from Speech. 3653-3657. 10.21437/Interspeech.2018-2222.

Cohen, A.S., Fedechko, T.L., Schwartz, E.K., Le, T.P., Foltz, P.W., Bernstein, J., Cheng, J., Holmlund, T.B., and Elvevåg, B. (2019). Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. Journal of Abnormal Psychology, 128(2), pp. 97–105. https://doi.org/10.1037/abn0000397.

Cohen, A.S., Mitchell, K.R., Docherty, N.M., and Horan W.P. (2016a). Vocal expression in schizophrenia: Less than meets the ear. Journal of Abnormal Psychology; 125(2):299-309. doi: 10.1037/abn0000136.

Cohen, A.S., Renshaw, T.L., Mitchell, K.R., and Kim, Y. (2016b). A psychometric investigation of "macroscopic" speech measures for clinical and psychological science. Behavior Research Methods. 48(2):475-86. doi: 10.3758/s13428-015-0584-1.

Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In Proceedings of the 2nd ACL Workshop on Computational Linguistics and Clinical Psychology.

Cowan, T., Le, T.P., Elvevåg, B., Foltz, P.W., Tucker, R.P., Holmlund, T.B., Cohen, A.S. (2019). Comparing static and dynamic predictors of risk for hostility in serious mental illness: Preliminary findings. Schizophrenia Research. 204, 432-433. doi: 10.1016/j.schres.2018.08.030

Cowie, R., McKeown, G., and Douglas-Cowie, E. (2012). Tracing emotion: an overview. Int. J. Synth. Emot. 3, 1–17. doi: 10.4018/jse.2012010101

Elvevåg, B., Foltz, P.W., Weinberger, D.R. and Goldberg, T.E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophrenia Research 93(1-3): pp. 304-316.

Eyben, F., Weninger, F. Gross, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In Proceedings of the 21st ACM International Conference on Multimedia, pp. 835–838.

Goodkind, A., Lee, M., Martin, G.E., Losh, M., and Bicknell, K. (2018). Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. In Proceedings of the Society for Computation in Linguistics, volume 1.

Grimm, M., Kroschel, K., and Narayanan, S.S. (2007). Support vector regression for automatic recognition of

spontaneous emotions in speech. In Proceedings of ICASSP, pp. 1085-1088.

Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A.S., D., Cheng, J., Bernstein, J., Rosenfeld, E., and Elvevåg, B. (in press / 2020a). Applying speech technologies to assess verbal memory in patients with serious mental illness. npj Digital Medicine.

Holmlund, T.B., Fedechko, T.L., Elvevåg, B. & Cohen, A.S. (2020b). Chapter 28: Tracking language in real time in psychosis. In: A Clinical Introduction to Psychosis: Foundations for Clinical and Neuropsychologists. Eds. J.C. Badcock & G. Paulik-White. Elsevier.

Holmlund, T.B., Foltz, P.W., Cohen, A.S., Johansen, H.D., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E. & Elvevåg, B. (2019). Moving psychological assessment out of the controlled laboratory setting and into the hands of the individual: Practical challenges. Psychological Assessment. 31(3), 292-303. doi: 10.1037/pas0000647

Landauer, T.K., Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, vol. 104, no. 2, pp. 211–240.

Le, T.P., Cowan, T., Schwartz, E.K., Elvevåg, B., Holmlund, T.B., Foltz, P.W., Barkus, E. & Cohen, A.S. (2019). The importance of loneliness in psychotic-like symptoms: Data from three studies. Psychiatry Research. doi: 10.1016/j.psychres.2019.112625.

Le, T.P., Elvevåg, B., Foltz, P.W., Holmlund, T.B., Schwartz, E.K., Cowan, T, Cohen, A.S. (2018). Aggressive urges in schizotypy: Preliminary data from an ambulatory study. Schizophrenia Research, 201, 424-425. doi: 10.1016/j.schres.2018.05.045

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.

Mower, E., Metallinou, A., Lee, C., Kazemzadeh, A., Busso, C., Lee, S., et al. (2009). "Interpreting ambiguous emotional expressions," in The 3rd International Conference on Affective Computing and Intelligent Interaction (Amsterdam), 1–8.

Pennebaker, J.W., Boyd, R.L., Jordan, K. and Blackburn, K. (2015). The development and psychometric properties of LIWC. Austin, TX: University of Texas at Austin.

Peters, M.E., Neumann, M., Iyyer, M., Gardne,r M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep contextualized word representations. NAACL-HLT.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F. Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In Interspeech, pp. 148–152.

Schuller, B.W., Steidl, S., Batliner, A., Marschik, P.B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F., Rathner, E.M., Bartl-Pokorny, K.D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., Tzirakis, P., and Zafeiriou, S. (2018). "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats," in INTERSPEECH, pp. 1–5.

Steidl, S., Levit, M., Batliner, A., Noth, E., and Niemann, H. (2005). "Of all things the measure is Man' automatic classification of emotions and inter-labeler consistency," in ICASSP (Philadelphia, PA), pp 317–320.

Sun, Y., Ayaz, H., and Akansu, A.N. (2020). Multimodal Affective State Assessment Using fNIRS + EEG and Spontaneous Facial Expression. Brain Sci. 2020;10(2):E85. doi:10.3390/brainsci10020085

Voleti, R., Liss, J.M., and Berisha, V. (2020). A Review of Language and Speech Features for Cognitive-Linguistic Assessment. in IEEE Journal of Selected Topics in Signal Processing - Special Issue on Automatic Assessment of Health Disorders Based on Voice, Speech and Language Processing.

Watson, D., Clark, L.A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology, vol. 54, pp. 1063–1070.

Zhaocheng, H. and Epps, J. (2018). Prediction of Emotion Change from Speech. Frontiers in ICT, vol. 5, pp. 11. doi: 10.3389/fict.2018.00011