2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Depression Classification Using *n*-Gram Speech Errors from Manual and Automatic Stroop Color Test Transcripts

Brian Stasak, Zhaocheng Huang, Julien Epps, and Dale Joachim

Abstract— **While the psychological Stroop color test has frequently been used to analyze response delays in temporal cognitive processing, minimal research has examined incorrect/correct verbal test response pattern differences exhibited in healthy control and clinically depressed populations. Further, the development of speech error features with an emphasis on sequential Stroop test responses has been unexplored for automatic depression classification. In this study which uses speech recorded via a smart device, an analysis of *n*-gram error sequence distributions shows that participants with clinical depression produce more Stroop color test errors, especially sequential errors, than the healthy controls. By utilizing *n*-gram error features derived from multi-session manual transcripts, experimentation shows that trigram error features generate up to 95% depression classification accuracy, whereas an acoustic feature baseline achieve only upwards of 75%. Moreover, *n*-gram error features using ASR transcripts produced up to 90% depression classification accuracy.**

I. Introduction

Cognitive interference tasks help to study the effects of mental health disorders. A commonly used cognitive load task is the Stroop color test [1, 2], which was developed to study human behavior and cognitive bias. Studies [3-6] have used on the Stroop color test to investigate depression and the impact of this mood disorder on cognitive response. When compared to healthy controls during Stroop color tests, individuals with depression have demonstrated a deterioration of cognitive function, including psychomotor slowness and more difficulty ignoring irrelevant information [3, 4].

The degree of error patterns made during the Stroop color test have received little attention. For example, in [3, 4], no significant differences in correct/incorrect responses between healthy and depressed individuals were found. However, [3] only examined whether the entire Stroop color test session was correctly or incorrectly completed, and not the correctness of individual color-text items. Furthermore, while [4] examined the correctness of individual color-text items per Stroop color test session, they did not evaluate successive response patterns. Remarkably, in the Stroop color test literature, we could not find any study that purposefully investigated sequential verbal error patterns in healthy and depressed individuals - or for that matter, other types of mental illness. However, studies [7-9] examining read aloud passages and spontaneous speech have indicated that individuals with depression exhibit greater frequency of cognitive impairment, malapropisms, referential failure, and verbal disfluency than healthy controls.

In [10], patients with depression revealed an abnormal behavioral response to their poor performance during CANTAB battery cognitive tasks. In a depressed population relative to a healthy control, it was shown that failure on task problem the first time increased the odds of failure on the next problem. According to [10], this unusual behavioral response (e.g., increased task errors, sequential errors) and its ‘snowballing effect’ may be a key indicator of depression severity.

Automatic techniques using speech processing have been explored to help identify individuals with depression. The prevailing method used to identify depression from the speech signal has included acoustic-based features (e.g., glottal, prosodic, spectral); however, linguistic-based features (e.g., type-tokens, syntax) derived from speech transcripts have also shown effectiveness [11]. Recent studies [11, 12] on automatic speech-based depression classification have advocated for speech-to-text feature techniques on account of its relatively low feature dimensionality. In [12], it was advised that an examination beyond acoustic properties will help yield new features that capture syntactic structures (e.g., word sequences) unique to individuals with and without depression.

The research in this paper is motivated by the shortage of analysis of error types in previous speech-based Stroop color test mental health literature [2-4]. Furthermore, semi-grounded on [10, 12], we wish to investigate the crucial link between a person’s mood state and his/her task performance, especially with consideration towards serial cognitive-verbal tasks. Due to cognitive-motor impairment, a sub-symptom of depression, we hypothesize that individuals suffering from depression will exhibit a greater number of errors during the Stroop color test than a healthy population. Also, based on the knowledge that as mistakes are made (i.e., depressive individuals abnormally exhibit less motivation to improve their task performance when compared to a healthy population [10]) we hypothesize that individuals with depression will exhibit greater frequency of consecutive Stroop color test errors than healthy controls.

II. Database

The speech recordings used in this study were privately collected in the Netherlands and consisted of 10 non-depressed healthy control (HC) and 10 clinically depressed English-speaking participants (CD) (see Fig. 1). For both the

*Research supported by ARC Linkage Project (LP160101360).

Brian Stasak is with the School of Elect. Eng. & Telecomm., University of New South Wales, Sydney, NSW AUS (e-mail: b.stasak@unsw.edu.au).

Zhaocheng Huang was with School of Elect. Eng. & Telecomm., University of New South Wales, Sydney, NSW, AUS. He is now with Amazon, Seattle, WA USA (e-mail: zhaocheng.huang@unsw.edu.au).

Julien Epps is with the School of Elect. Eng. & Telecomm., University of New South Wales, Sydney, NSW, AUS (e-mail: j.epps@unsw.edu.au).

Dale Joachim is with Sonde Health, Boston, MA, USA (e-mail: djoachim@sondehealth.com)

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

HC and CD, the average age of a participant was 30 years old with a similar age range of 19 to 53 years old. The dataset described in this paper was approved by the IRB.

All participants were evaluated by a mental health expert using a structured interview and given the Montgomery-Åsberg Depression Rating Scale (MADRS) [13]. The MADRS is a common ten-item diagnostic questionnaire used to help evaluate depression disorder severity. The MADRS has five depression severity label score ranges: *normal* (0-6), *mild* (7-19), *moderate* (20-34), and *severe* (35-60).

Using a smart device app, all participants were asked to complete a series of ten Stroop color test sessions in English, with each session consisting of ten color-text items (i.e., 10 sessions × 10 color-text items). There were ten possible color-text items (e.g., black, blue, brown, green, gray, pink, purple, red, orange, yellow). Per item, the color-text mismatch or match order was randomized for every speaker.

Sessions were in a quiet office over the course of roughly a single day. For each session, a series of ten individual color-text words were displayed in 800ms intervals. During each Stroop color test session, participants were instructed to say the ink color of the printed word and not the actual word. The average participant file length per session was 10sec.

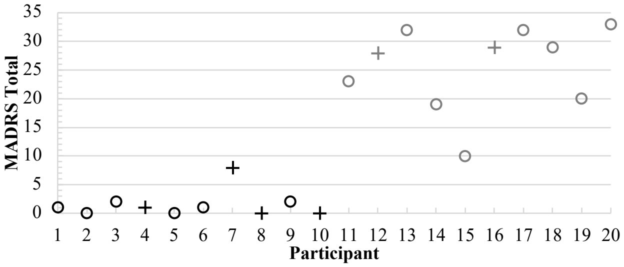

Figure 1. Stroop color test database female (°) and male (+) participant MADRS severity score distributions. Non-depressed participants are indicated in black (≤7 MADRS), whereas clinically depressed participants are in gray (≥10 MADRS).

III. Methods

A. Human vs. Automatic Transcripts

All participant speech recordings were transcribed by a native English-speaking annotator with a background in speech science. The annotator was instructed to focus on the verbal sequence of color words related to the Stroop color test. Therefore, the annotator ignored any extraneous verbal participant interjections (e.g., '*uh*', stammers, cusses). Also, if participants made an attempt to change their verbal response, their revisions were lodged in the transcript rather than their initial response. Approximately 5% of the recordings contained spoken terms outside of the ten possible color words, especially in recorded sessions where errors were more frequent.

Each participant's recorded Stroop color test session was also processed using Amazon Web Services™ Automatic Speech Recognition (ASR) software using a large vocabulary Dutch-English language model. To correct minor ASR transcription errors a python script with a similar word look-up list was executed to replace homonyms (e.g., read/red, blue/blew) and similar sounding terms (e.g., black/back, red/ready, yellow/hello). This script also ignored common insertions, such as '*uh*' or '*um*'. Measures of phonetic similarity (e.g., edit distance), have been implemented previously with ASR systems to help assess transcript errors [14].

B. *n*-Gram Error Feature Set

This section describes how the proposed *n*-gram error transcript-based features were calculated. Each session has a ground truth transcript (i.e., *ink color of word*) $\boldsymbol{w}_s = \{w_{1,s}, \dots w_{k,s}, \dots, w_{K,s}\}$ and a spoken transcript (i.e. *what color participant actually said*) $\widehat{\boldsymbol{w}}_s = \{\widehat{w}_{1,s}, \dots \widehat{w}_{k,s}, \dots, \widehat{w}_{K,s}\}$, where $w_{k,s}$ represents the k^{th} word in the s^{th} session. A correct response '1' is recorded if $\widehat{w}_{k,s}$ matches $w_{k,s}$, otherwise it is an incorrect response '0'.

$$c_{k,s} = \begin{cases} 0 & \widehat{w}_{k,s} \neq w_{k,s} \\ 1 & \widehat{w}_{k,s} = w_{k,s} \end{cases} \tag{1}$$

Based on (1), the transcript $\widehat{w}_s$ can be converted into a set of binary strings indicating the correctness of responses $\boldsymbol{c}_s$=$\{c_{1,s}, \dots, c_{k,s}, \dots, c_{K,s}\}$, e.g., if $\boldsymbol{w}_s = [r, o, y, pi, bl, r, gr, y, r, o]$ and $\widehat{\boldsymbol{w}}_s = [r, o, bl, pi, gr, r, gr, y, r, r]$, then $\boldsymbol{c}_s = [1,1,0,1,0,1,1,1,1,0]$.

Based on $\boldsymbol{c}_s$, we consider unique c_s^i, and sequential patterns $c_s^{i,j}$ and $c_s^{i,j,m}$, which are referred to as unigrams, bigrams and trigrams, where $i, j, m \in \{0, 1\}$. The number of occurrences was then calculated for all possible patterns c_s^i, $c_s^{i,j}$ and $c_s^{i,j,m}$, leading to $\boldsymbol{c}_s^{n=1}$, $\boldsymbol{c}_s^{n=2}$ and $\boldsymbol{c}_s^{n=3}$ respectively, which were then concatenated to form the error-based pattern features $\boldsymbol{e}_s$.

$$\boldsymbol{e}_s = [(\boldsymbol{c}_s^{n=1})^T \ (\boldsymbol{c}_s^{n=2})^T \ (\boldsymbol{c}_s^{n=3})^T]^T \tag{2}$$

where $\boldsymbol{e}_s^{n=1} = [\#(c_s^0), \#(c_s^1)]^T$, $\boldsymbol{e}_s^{n=2} = [\#(c_s^{00}), \#(c_s^{01}), \#(c_s^{10}), \#(c_s^{11})]^T$, and $\boldsymbol{e}_s^{n=3} = [\#(c_s^{000}), \#(c_s^{001}), \#(c_s^{010}), \#(c_s^{100}), \#(c_s^{011}), \#(c_s^{101}), \#(c_s^{110}), \#(c_s^{111})]^T$, and # denotes the counting operation.

Effectively, the *n*-grams represent a distribution of pattern sequences per session that provide detailed information regarding the number of individual errors, number of errors in a row, and overall error pattern distributions. Once all *n*-gram error distributions were calculated for each of the ten Stroop color test sessions, sets of unigram, bigram, and trigram distributions were averaged per individual session to create a set of fourteen *n*-gram error features per participant. Consecutive sessions (e.g., 1-2, 1-3, …, 1-10) were averaged to create multi-session *n*-gram error feature sets.

C. Acoustic Feature Set

For an acoustic feature baseline, frame-level voice activity detection was applied to extract segments of speech only and frames that included silence were removed. Spectral features (e.g., MFCC, formants) were then extracted from the speech segments of each file using a 20ms window with 50% overlap. The 100-dimensional acoustic feature set included computed functionals for the mean, standard deviation, skewness, kurtosis, and 10%/90% percentiles.

D. System Configuration

Fig. 2 summarizes the experimental design steps. Initially, speech transcripts ($\widehat{w}_s$) per session were compared with the Stroop color test ground-truth (w_s) and converted into a binary string form (e_s). Binary string representations were then analyzed using *n*-gram counts per participant test session.

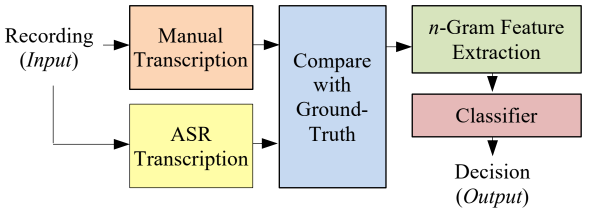

Figure 2. Experimental design for analyzing *manual* and *ASR* based *n*-gram errors derived from participants' verbally recorded responses of the Stroop color test.

E. Experimental Settings

Similarly to other speech-based depression studies [15, 16], due to the equal number of HC and CD participants, binary HC/CD classification accuracy was reported for experiments herein. This accuracy was computed by calculating the ratio of number of correct test identifications to the total number of test files per leave-one-out cross fold validation experiment, and then averaging all individual leave-one-out cross fold validation results. Therefore, no participant was found in both training and test during fold experiments. As a backend classifier, a linear support vector machine (SVM) with a polynomial kernel function (*order* = 3) was applied due to its robustness to overfitting and previous application in mental health studies [15-17]. The SVM used 3-fold cross validation to determine the optimal *C* parameter (e.g., ranged from log 10^{e-5} to 10^{e1}).

While this experimental database consists of multi-session data, it contains a relatively small participant sample size resulting in potentially low statistical power. Unfortunately, the sensitive nature of mental illness and patient privacy limits access to speech-based clinical depression data in general. For example, there are currently no publicly available recordings of individuals with depression completing the Stroop color test.

IV. Results and Discussion

Over many sessions, due to the Stroop color test cognitive processing interference, it was anticipated that even for the HC participants, sporadic errors would be recorded. Based on manual transcripts, the HC participants completed 67% of the test sessions without producing a single mistake on a test item, whereas the CD participants achieved only 56% of the test sessions error-free. Unlike previous studies [3, 4], analysis herein indicated that the CD participants had more difficulty with avoiding an error during multi-session Stroop color tests.

Further analysis based on manual transcripts shown in Fig. 3 indicated that the CD participants produced more *n*-gram '0' (i.e., errors) than the HC. For example, using all test sessions, the average number of '0' was 2.06 for the CD participants, whereas it was only 1.39 for the HC participants. Furthermore, as anticipated, CD participants had a higher average for sequential *n*-gram errors (e.g., '00', '000', '001', '100') than the HC participants.

Interestingly, manual transcript *n*-gram error patterns shown in Fig. 3 also demonstrated that the HC participants had an approximately 30% increase in average '101' occurrences than the CD participants. Therefore, to a moderate degree, after an error, the HC participants are more capable of recovering from a mishap on their next response than the CD participants according to the manual transcripts. The CD participants' inability to recover after an error was further supported as on average they had '001' and '100' occur more than twice as often as the HC participants.

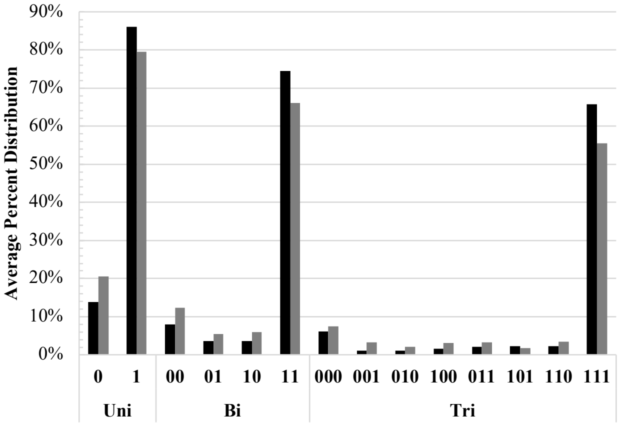

Figure 3. Average percentages of *n*-gram errors over sessions 1-10 based on *manual* transcripts for HC (black) and CD (gray) participants. The '0' represents an incorrect Stroop color word response, whereas a '1' represents a correct response.

Since manual transcription is labor intensive, a natural question is how well *n*-gram features perform using a transcript generated from a fully automated method. In Fig. 4, average *n*-gram analysis based on ASR transcripts also showed that CD participants demonstrated an increase in errors when compared with HC participants. For the ASR-based transcripts, a slight increase in '0' *n*-gram errors (8%-13%) when compared with the manual transcripts was observed for both HC and CD participants. This increase was attributed to automated transcript errors (i.e., higher word-error rate than manual method) and difficulty automatically assessing participants' revisions. Nevertheless, in comparing the average *n*-gram distributions derived from manual and ASR transcripts, similar distribution trends were recorded (i.e., 12 out of 14). For example, for both the manual and ASR based transcripts, sequential *n*-gram errors (e.g., '00', '000', '001', '100') had higher values for the CD than the HC participants.

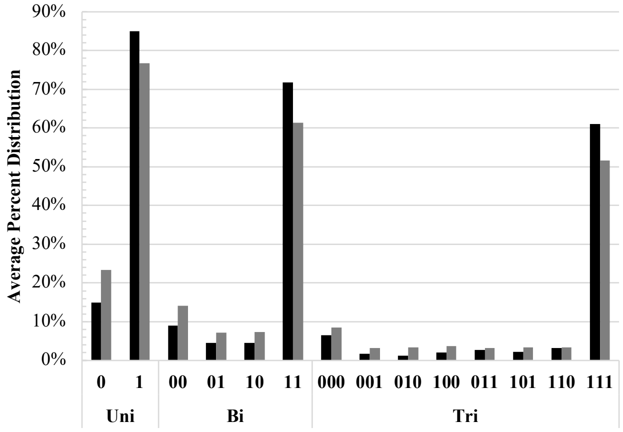

Figure 4. Average percentages of *n*-gram errors for sessions 1-10 based on *ASR* transcripts for HC (black) and CD (gray) participants.

Shown in Fig. 5, the acoustic, unigram, bigram, trigram, all *n*-grams, and trigram best features were evaluated for depression classification. The trigram best features were based on trigrams that demonstrated the most separability between HC and CD classes found previously in Fig. 3 and 4. When compared with trigram error features, classification accuracy results indicated that acoustic, unigram and bigram feature sets were less accurate at classifying participants. It is believed that the trigram feature set (8 dims.) did well because it consists of more unique sequential error pattern information, whereas the unigram (2 dims.) and bigram (4 dims.) feature sets were more limited.

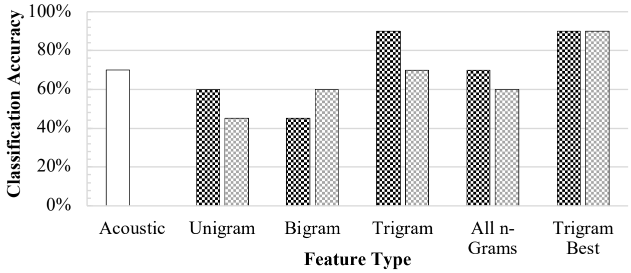

Figure 5. 2-class (HC/CD) classification results using leave-one-out cross validation for acoustic (solid white), *n*-gram *manual* (black pattern) and *ASR* (gray pattern) transcript *n*-gram error feature sets based on all sessions combined (1-10).

To a small degree, the manual transcript features generally outperformed the ASR based features (6% average increase); due in part to the increased word-error rate found in the ASR transcripts. Using optimal 'best' trigram feature selection, based on analysis of Fig. 3 and 4 (e.g., '001', '010', '101', '110'), depression classification results were maximized (90%) for manual and ASR derived features.

An investigation of feature type and accuracy per individual session was also conducted. In Fig. 6, the first session produced the lowest depression classification accuracies for the *n*-gram error features. This weaker first session depression classification result when compared to subsequent sessions, is attributed to initial test learning phenomena; wherein participant task familiarity, expectation, and ability increases as more sessions are completed.

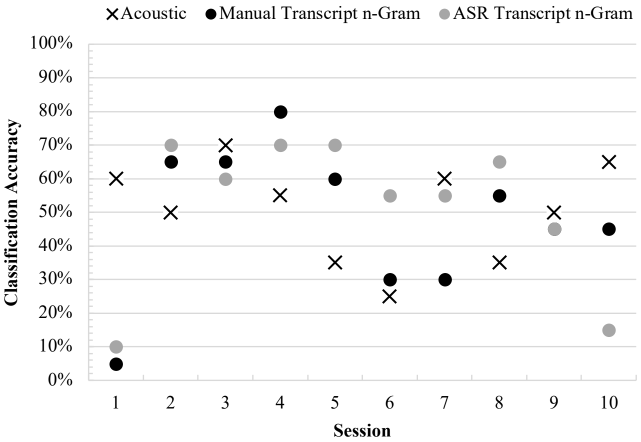

Figure 6. Individual Stroop color test 2-class (HC/CD) classification results using acoustic feature set (100 dims.), all *manual n*-gram error feature set (14 dims.) and all *ASR n*-gram feature set (14 dims.) transcripts.

For future Stroop color test collections, it is advised that a fair degree of participant practice is allotted; or that the initial test session be precluded from analysis, as it is generally less dependable than later sessions. Results shown previously in Fig. 6, and also in Fig. 7 and 8, demonstrated classification accuracy variance was greatest during the initial Stroop color test sessions 1-3.

In addition to individual session feature depression classification, experiments were conducted to investigate features from accumulated sessions. Across multiple successive test sessions, the acoustic features maintained a depression classification accuracy range of 55%-75%, whereas the *n*-gram error feature set had a range of 5%-85%. However, it is shown in Fig. 7 that as the number of successive test sessions were averaged into the *n*-gram error feature set, stability and improvements in its depression classification accuracy were reported. For both manual and ASR derived features, computing *n*-gram error features from ≥7 sessions information results in a relatively less variable depression classification accuracy than using ≤4 sessions. Further, a manual *n*-gram error feature experiment was conducted using consecutive sessions in reverse (e.g., 10, 10-9, 10-8, …, 10-1), wherein depression classification gains were again observed beyond ≥7 sessions with up to 85% depression classification accuracy.

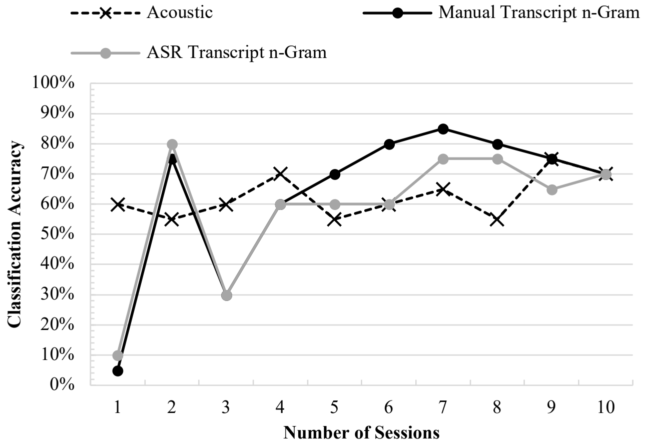

Figure 7. 2-class (HC/CD) classification results showing a comparison of acoustic (100 dims.), all *manual* (14 dims.) and *ASR* (14 dims.) transcript-based *n*-gram error features.

Shown in Fig. 8, depression classification results using the trigram features with ≥9 sessions produced higher accuracy than using all *n*-gram error features (see previous Fig. 7). Again, it was observed that as the number of sessions increased, so did depression classification accuracy. Furthermore, although not shown, using manual transcript four best trigram features (e.g., '001', '010', '101', '110') the highest depression classification accuracy (95%) was recorded using sessions 1-8 and 1-9. Again, similarly to other feature types, utilizing ≥7 sessions led to improved depression classification accuracy for the best trigram error feature set.

The steady rise in depression classification performance as more successive Stroop color test sessions were included in the *n*-gram error feature sets shown in Fig. 7 and 8 were attributed to the CD participants propensity for motivational fatigue and negative response to failure [10]. For manual transcript '0' (e.g., errors), the ND participants averaged 2.02 in sessions 1-5 and 0.78 in sessions 6-10; whereas the CD participants averaged 1.85 in sessions 1-5 and 1.83 in sessions 6-10. With additional session practice, the ND participants were able to reduce their average number of errors by more than half. However, the CD participants maintained approximately the same average number of errors as sessions increased.

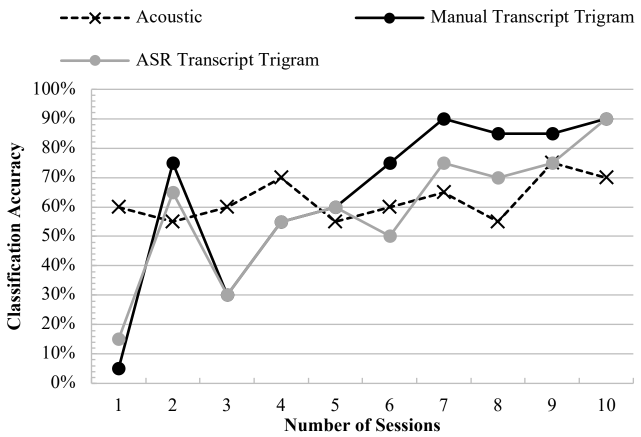

Figure 8. 2-class (HC/CD) classification results for *manual* (8 dims.) and *ASR* (8 dims.) transcript-based trigram error features.

V. Conclusion

In this study, we proposed a novel approach to speech-based depression classification using a binary analysis of Stroop color test responses and *n*-gram error features. We found that participants with depression demonstrated a higher frequency of errors during the Stroop color test than the healthy control. Moreover, we revealed a crucial link between clinically depressed participants' mental health status and their impaired ability to recuperate with correct responses after Stroop color test errors than the healthy control. Depression classification accuracy results using low dimensional *n*-gram error features were competitive with the baseline acoustic features. Further, as the number of sessions used to compute *n*-gram features increased depression classification accuracy improved. This study shows that with sufficient Stroop color test session data (≥7 sessions), low-dimensional manual or ASR-derived trigram error features can produce relatively good depression classification accuracy (75%-95%).

Based on our experimental Stroop color test results, more research on *n*-gram error features is warranted. It is believed that the Stroop color test *n*-gram error features may also prove useful for identifying or monitoring other types of neurological illnesses, such as: dementia, traumatic brain injury, and multiple sclerosis.

References

[1] R. Stroop, Studies of interference in serial verbal reactions, *J. Experimental Psych*., vol. 18, pp. 643–662, 1935.

[2] F. Scarpina and S. Tagini, The Stroop color and word test, *Front Psycho*l., vol. 8, pp. 557, 2017.

[3] B. Gohier, L. Ferracci, S.A. Surguladze, E. Lawerance, W. El Hage, M.Z. Kefi, P. Allain, J.B. Garre, and D. Le Gall, Cognitive inhibition and working memory in unipolar depression, *J. Affect. Disord*., vol. 116, no. 1-2, pp. 100–109, 2009.

[4] S. Kertzman, I. Reznik, T. Hornik-Lurie, and A. Weizman, Stroop performance in major depression: selective attention impairment or psychomotor slowness?, *J. Affect. Disord*., vol. 122, no. 1-2, pp. 167–173, 2009.

[5] M.T. Mitterschiffthaler, S.C. Williams, N.D. Walsh, A.J. Cleare, C. Donaldson, J. Scott, and C.H. Fu, Neural basis of the emotional Stroop interference effect in major depression, *Psychol. Med*., vol. 28, no. 2, pp. 247–256, 2008.

[6] A.M. Epp, K.S. Dobson, D.J. Dozois, and P.A. Frewen, A systematic meta-analysis of the Stroop task in depression, *Clin. Psychol. Rev*., vol. 32, no. 4, pp. 316–328, 2008.

[7] S.M. Levens and I.H. Gotlib, Updating emotional content in recovering depressed individuals: evaluating deficits in emotion processing following a depressive episode, *J. Behav. Ther. Exp. Psych*., vol. 48, pp. 156–163, 2015.

[8] I.A. Rubino, L. D'Agostino, L. Sarchiola, D. Romeo, A. Siracusano, and N.M. Docherty, Referential failures and affect reactivity of language in schizophrenia and unipolar depression, *Schizophrenia Bulletin*, vol. 3, no. 3, pp. 554–560, 2011.

[9] B. Stasak and J. Epps, Automatic depression classification based on affective read sentences: opportunities for text-dependent analysis, *Speech Comm*., vol. 115, pp. 1–14, 2019.

[10] R. Elliot, B.J. Sahakian, J.J. Harrod, T.W. Robbins, and E.S. Paykel, Abnormal response to negative feedback in unipolar depression: evidence for a diagnostic specific impairment, J. *Neurology, Neurosurgery, & Psychiatry*, vol. 63, no. 1, pp. 74–82, 1997.

[11] B. Stasak, J. Epps, and R. Goecke, Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect, In: *Proc. Of InterSpeech, Stockholm,* Sweden, pp. 834–838, 2017.

[12] M.R. Morales and R. Levitan, Speech vs. text: a comparative analysis of features for depression detection systems, In: *Proc. IEEE Workshop on Spoken Language Technology*, San Diego, CA, pp. 136–143, 2016.

[13] J. Davidson, C.D. Turnbull, R. Strickland, R. Miller, and K. Graves, The Montgomery-Åsberg depression scale: reliability and validity, Acta Psych. Scand., vol. 73, no. 5, pp. 544–548, 1986.

[14] R. Errattahi, A.E. Hannani, and H. Ouahmane, Automatic speech recognition errors detection and correction: a review, In: Proc. Intern. Conf. on Natural Lang. and Speech Process. (ICNLSP), *Procedia Computer Science*, vol. 128, pp. 32–37, 2015.

[15] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, An investigation of depressed speech detection: features and normalization, In: *Proc. Annu. Conf. Int. Speech Commun. Assoc*., pp. 2997–3000, 2011.

[16] H. Jiang, B. Hu, Z. Riu, G. Wang, L. Zhang, X. Li, and H. Kang, Detecting depression using ensemble logistic regression model based on multiple speech features, *Comput. Math. Methods Med*., vol .2018, pp. 1–9, 2018.

[17] G. Orrú, W. Pettersson-Yeo, A.F., Marquand, G. Sartori, and A. Mechelli, Using a support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review, *Neurosci. & Biobehav. Reviews*, vol. 36, no. 4, pp. 1140–1152, 2012.