Read Speech Protocol Criteria for Speech-Based Health Screening Applications

Brian Stasak[1,2] *and Julien Epps*[2]

[1]Black Dog Institute, Sydney, NSW – Australia
[2]School of Elec. Eng. & Telecom., UNSW Sydney, NSW – Australia
b.stasak@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Automatic speech-based processing using machine learning is expanding in digital healthcare, bolstering potential as a non-invasive, remote medical screening tool. There is currently a need for deeper understanding of read speech protocols and applying systematic measurements to help tailor new protocols with deliberate attribute criteria. This study investigates eight read speech protocols commonly utilized to study speech behaviours and proposes two new protocols with greater criteria extremes. An investigation of text, phonetic, linguistic, and affective proposed criteria automatically extracted from these read speech protocols reveals important merits and limitations for use in speech-based digital health screening applications.

Index Terms: data collection; elicitation; voice processing

1. Introduction

Healthcare clinicians perform speech-language evaluations to screen, diagnose, and monitor many disorders [1, 2]. During a structured interview, clinicians observe the patient's speech production, such as articulation, breathing, phonation, and voice quality. Clinicians also evaluate a patient's spoken language ability, including grammar, pragmatics, memory, and expressive capacity. Abnormal speech-language symptoms are often early precursors to a variety of disorders and illnesses.

Some clinical evaluations also include an analysis of speech behaviours based on read protocols, which consists of pre-selected words, sentences, or paragraphs that are read aloud. Advantages to read speech protocols include minimal instruction, repeatability, 'ground-truth' knowledge, limited cognitive scope, and controlled phonetic variability. Further, read speech protocols are relatively easy to implement in digital smart device apps. Among the most popular read speech protocols for speech-based analysis are: 'Arthur' [3], 'The Caterpillar' [4], 'The Farm Script' [5], 'Hunter Script' [5], 'The Grandfather Passage' [6], 'The John Passage' [7], 'The North Wind and the Sun' [8], and 'The Rainbow Passage' [9].[1]

When selecting an ideal read speech protocol for pathological medical analysis there are important factors, such as speaker background (e.g., age, reading skill level), illness specificity (i.e., focus on elicitation of key symptoms), and task duration (i.e., time duration, number of samples needed). Despite read speech protocol groundwork [6, 9-12], still many speech-based health screening studies [13-15] continue to use semi-antiquated read speech protocols (i.e., the origin of 'Arthur', 'The Grandfather Passage', and 'The North Wind and the Sun' were derived from writings more than a century old), which were not originally intended for 'universal' illness/disorder screening. Of the protocols mentioned, many were originally used to subjectively judge verbal intelligibility oral reading rates of school-aged children and not designed for precision assessment of a wide variety of different types of disorders (e.g., psychogenic, neurological, respiratory, voice) [3, 6, 8, 9]. For example, 'The Rainbow Passage' is nearly 80 years old, and it was based on child voice articulation drills described for '*correction of disorders*'; although it does not define which ones or provide statistical normative data [9].

While it is frequently believed that these read speech protocols contain every English phoneme and are phonetically balanced [16], previous studies [11, 17, 18] have shown many of these read speech protocols have non-ideal, disproportionate phoneme distribution ratios. Moreover, some of these protocols do not purposefully isolate or repeat specific phonemes, phonetic transitions, and/or language components that may be more useful in more direct screening for certain illnesses/disorders. Increasing the opportunity for phonemes in read speech protocols that ordinarily are less frequently found in conversational speech may be advantageous, especially if a disorder (e.g., aphasia, apraxia, dysphonia) affects production of a specific phoneme or phoneme class.

From an age-appropriate readability standpoint, many of these read speech protocols [3, 7, 8, 9] are third-person narratives that include unfamiliar themes and advanced vocabulary (i.e., 'The North Wind and the Sun' is a translation based on Aesop's ancient Greek fable [8]). This may add more difficulty reading aloud when compared with natural conversational speech. Also, some of these read speech protocols include unsuitable themes for younger children (i.e., 'Arthur' has a death; 'The Hunter' has a firearm).

Previous studies [4, 16-18, 20] have comparatively examined read speech protocols. However, these previous studies focused primarily on text-based attributes or phonetic distributions rather than deeper level structure criteria, such as age of articulatory mastery and gestural phonetic transition effort. Only two of these previous studies [4, 20] contributed new read speech protocols – working to further expand the acoustic speech elicitation space.

Automatic digital speech-based screening studies [4, 13, 21-26] involving the aforementioned protocols typically comprised a relatively limited number of speakers (i.e., less than two dozen) and each contained dissimilar speakers (e.g., children, adults, elderly). Also, these automatic speech-based illness screening studies examined the effectiveness of just one or two read speech protocols. Further, only recently have speech-based illness detection studies [14, 27-30] examined linguistic/affect components in read speech protocols. For example, Boaz et al. [27] investigated 'The Rainbow Passage'

[1] Free access to the speech protocols in this study are available at: https://www.researchgate.net/profile/Brian_Stasak/projects/

and how its affective content impacted speech disfluencies. They found ‘The Rainbow Passage’ and a novel constructed passage disfluency types (e.g., pauses, repeats) were not significantly different. However, half of the individuals studied showed a large difference in the number of disfluencies between the two passages, likely due to readability factors.

This paper investigates text, phonological, linguistic, and affective attributes from ten read speech protocols, including two new speech protocols. New articulatory criteria regarding age of acquisition mastery, gestural effort, and phoneme-to-word ratios are reported along with examples of how different criteria are relevant to disorders. Results herein show that some read speech protocols are more alike than others. This deeper level analysis adds to the discussion on the need for modern read speech protocols with more methodical design criteria, including phonetic, linguistic, and affect factors, resulting in greater knowledge of how age- and skill-appropriate a protocol is for a particular speech-based medical screening.

2. Analysis and Discussion

2.1. Protocols and Criteria System

Eight existing read speech protocols were examined in this study (see Tables 1 and 2). These established protocols were selected for analysis due to their frequent use in previous speech-based studies [4, 13, 16-18, 20-26]. Two newly proposed read speech protocols, ‘Jazz’ and ‘Restaurant’, were created using deliberately chosen words to induce higher articulatory skill level and unique read token word demand.

In Figure 1, a system design for automatic extraction of criteria based on existing or new speech protocol texts is proposed. Using raw text as input, extraction of various feature types produces an output, whereby feature values are then compared to other read texts. This allows an understanding of deeper level information about what a read protocol contains. A set of feature criteria ranges can be experimented with to help better tailor read speech protocols for more specific precision disorder screening. This proposed system also can be utilized for automatic data selection of non-read free speech transcripts to extract single phrases with feature criteria ranges of interest.

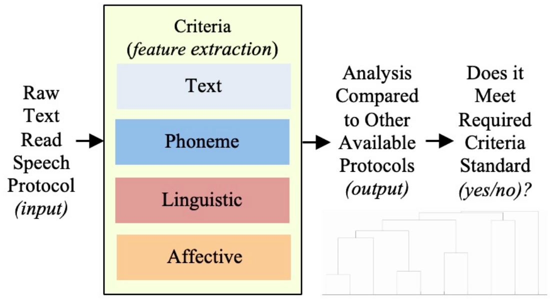

Figure 1: Proposed multi-dimensional automatic read speech protocol analysis design, which provides in-depth knowledge of different criteria, important when considering the appropriateness of read speech protocol for illness screening applications and test subject demographics. A hierarchical cluster tree method (e.g., Chebychev, Spearman) can be used to calculate read speech protocol similarity to other protocols [17].

2.2. Text and Phoneme Criteria

Surface-level text-based analysis provides limited articulatory insight. For example, a six-letter word like ‘*though*’ only contains two phonemes, whereas a six-letter word like ‘*plants*’ contains six unique phonemes. Therefore, text-based letter/word counts can be misleading in terms of articulatory ground-truth (e.g., *fizz*, *physics*, *laugh*), phoneme transition articulatory demand, and phoneme distributions.

The text-based analysis of the read speech protocols is shown in Table 1. Per protocol, the total number of words varied from as low as 97 to as high as 338. The mean length utterance (i.e., average number of words per sentence) of a read speech protocol, also referred to as MLU, may be an important consideration for automatic speech-based illness screening applications because recorded read aloud tasks that require minutes to complete require more time, storage, and user commitment/focus. Table 1 shows that the ‘The North Wind

Table 1: Read speech protocol comparison of basic text and phonetic information. This also includes never-before reported protocol scores for articulatory acquisition mastery and gestural effort measures [28, 29]. Newly proposed read speech protocols are shaded.

Protocol	# Words	# Unique Words	# Sent.	MLU	Word Range per Sent.	Aver. TTR	# Phonemes	MLP	Phon. # Range per Sentence	Aver. P/W Ratio	Aver. Art. Mast.	Aver. Gest. Eff.
Arthur	338	197	30	11.27	1 – 24	0.58	1023	33.1	4 – 74	3.12	62	6.3
Caterpillar	195	115	16	12.19	4 – 26	0.59	664	41.5	9 – 77	3.37	61	6.5
Farm	313	169	16	19.56	5 – 34	0.54	902	56.3	16 – 101	2.87	62	6.6
Grandfather	131	99	7	16.63	8 – 29	0.76	475	59.4	27 – 116	3.59	62	6.3
Hunter	279	151	17	16.41	4 – 35	0.54	878	51.7	13 – 109	3.21	62	6.4
John	191	119	11	17.36	7 – 29	0.62	617	56.0	25 – 91	3.25	58	6.8
North	113	64	5	22.60	14 – 36	0.57	384	76.8	52 – 115	3.44	62	6.5
Rainbow	330	175	19	17.37	8 – 36	0.53	1146	60.3	26 – 117	3.49	62	6.4
Jazz	117	88	12	9.75	5 – 16	0.75	478	39.8	21 – 64	4.15	67	6.3
Restaurant	97	76	9	10.78	5 – 14	0.78	446	49.6	25 – 60	4.68	65	6.7

Table 2: Read speech protocol comparison of basic linguistic and affective information. This includes Flesch-Kincaid grade-level and reading ease, Dirichlet allocation age of word exposure and acquisition scores [4, 31, 32]. Affect averages are also shown derived from Affective Norms for English Words (ANEW) [33]. All affective scores had a range from 1 (low) to 9 (high).

Protocol	% Passive Voice	Aver. Grade Level	Aver. Reading Ease	Aver. Age Word Exposure	Aver. Age Word Acquisition	Arousal	Dominance	Valence
Arthur	0.0	1.9	100	3.0	4.98	4.99	5.10	4.93
Caterpillar	0.0	5.0	81	3.7	5.07	5.33	5.29	6.26
Farm	18.7	2.6	100	2.5	4.86	4.72	5.51	6.22
Grandfather	0.0	5.2	81	3.5	5.77	4.89	5.66	6.53
Hunter	0.0	3.9	96	3.3	5.17	4.86	4.96	6.06
John	9.0	6.4	80	4.0	5.64	4.63	5.53	6.24
North	40.0	8.2	76	4.5	5.55	5.43	6.40	7.03
Rainbow	15.7	7.7	70	4.4	5.48	5.07	5.33	8.86
Jazz	8.3	6.6	64	6.4	6.58	5.33	5.40	6.52
Restaurant	0.0	9.5	46	5.0	6.30	5.03	5.69	6.66

and the Sun' protocol has the largest average for number of words per sentence (22.6), whereas the new 'Jazz' protocol averages much less (9.75).

The number of unique words was much less than the total number of words for all protocols (76–197). A type-token-ratio (TTR) is defined as the relationship between the number of unique words (e.g., core word types excluding word modifiers) and the number of total words (e.g., tokens). It is calculated by dividing the number of unique words by the total number of words. The more unique words there are in comparison to the number of total words, then the more lexical variety (i.e., a high TTR). Previously in Table 1, it was reported that the 'Restaurant' protocol had the highest TTR, whereas the 'The Rainbow Passage' demonstrated the lowest TTR.

Generally, the more words that are repeated, the more opportunity for like-word acoustic comparison. Therefore, depending on the illness/disorder being screened, it may be more effective to use a read speech protocol that has a low TTR to evaluate the same words more than once. On the contrary, if lexical diversity is of higher interest, it may be optimal to utilize a read speech protocol that has a high TTR. For example, for an early childhood speech sound disorder, having multiple examples of the same word may be good to determine the percentage of correct pronunciation, whereas for short-term memory disorders a read aloud memory test, a read speech protocol with more unique words helps, to increase cognitive demand and test specific keyword recall.

A phoneme-based analysis of the read speech protocols was conducted using thirty-nine English phonemes based on the Carnegie Mellon University phonetic dictionary [34]. A python script was created to convert the raw text of each word per read protocol into the most common phonetic representation. Results demonstrated that none of the read protocols were truly phonetically balanced; meaning they did not have equal representations for each English phoneme. Therefore, using these protocols, it may be difficult to analyse multiple examples of specific phonemes (e.g., /j, ʒ/) when compared with other phonemes that have a much greater frequency (e.g., /ə, n/). For example, utilizing a protocol with a higher frequency of rarer phonemes is important if close analysis of palatal or postalveolar positioned articulatory production is of high interest (e.g., speech sound disorders, palatal fronting).

In comparing each of the read speech protocols to standard norm English phoneme distributions found in natural free conversational speech [35], it was observed that the read speech protocols had a very strong 0.83-0.91 Spearman's rank correlation coefficient (a = 0.05). This analysis indicates that these read speech protocols contain similar phoneme distributions to natural speech. Phonemic analysis per read speech protocol showed large differences in the number of English phonemes, especially mean length of sentence phonemes (MLP) (i.e., the average number of phonemes per sentence). As shown previously in Table 1, although the 'Arthur' (1023) and 'The Rainbow Passage' (1146) contained the largest number of phonemes, 'The North Wind and the Sun' protocol had the largest average number of phonemes per sentence (~77) despite roughly one-third fewer sentences.

The phoneme-to-word ratio (P/W) is a better indicator of articulatory demand than using an average text-based word length or word total because it represents a ground-truth of spoken sounds. Further, read speech protocols with higher density phoneme-to-word ratios enable generation of more phonemes using fewer words – therefore, allowing increased efficiency in comparison to longer protocols that contain more words and recording time.

In terms of the phoneme-to-word ratio, which was calculated by dividing the number of phonemes by the total number of words, Table 1 and Figure 2 show that the 'Jazz' and 'Restaurant' new read speech protocols were much higher than the other traditional protocols. Analysis based on sentence-level average phoneme-to-word ratio scores indicated that the 'Jazz', 'Restaurant', and 'The Farm Script' were least like the other read speech protocols. The 'Hunter', 'John', 'Arthur', 'Caterpillar', and 'North Wind' were most alike in terms of phoneme-to-word ratio scores.

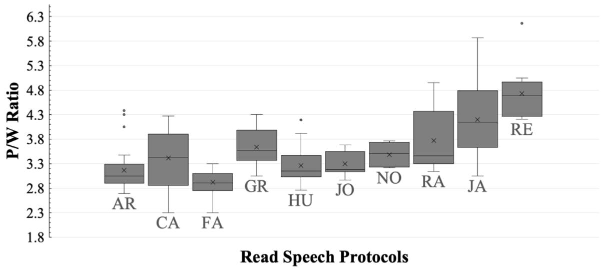

Figure 2: Sentence-level distributions of phoneme-to-word ratio (P/W) per read speech protocol: Arthur (AR), Caterpillar (CA), Farmer (FA), Grandfather (GR), Hunter (Hu), John (JO), North (NO), Rainbow (RA), Jazz (JA), and Restaurant (RE). Mean is indicated by 'x', the solid line indicates median, and outliers are represented as dots.

Previously, an acoustic speech-based study [28] showed that phrases containing a greater number of consonant phonemes mastered later in life were more effective in detecting individuals with depression. The reasoning is that phonemes mastered later in life require greater articulatory coordination, and are therefore a better measure for fine motor control. Shown in Figure 3, an assessment of age of articulatory acquisition mastery demonstrated that the 'Jazz' protocol had the highest average age in months (67), whereas the 'John' protocol had the lowest average (58). The new read speech protocols had an average articulatory age of acquisition mastery that was higher than the other existing protocols. The 'Arthur', 'Grandfather', 'Hunter', 'Rainbow', and 'North Wind' are most alike based on mean age of articulatory acquisition mastery.

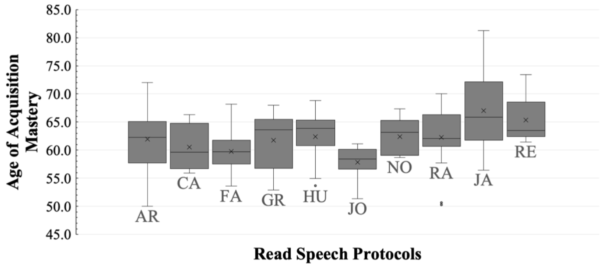

Figure 3: Sentence-level distributions of age of acquisition mastery per read speech protocol: Arthur (AR), Caterpillar (CA), Farmer (FA), Grandfather (GR), Hunter (HU), John (JO), North (NO), Rainbow (RA), Jazz (JA), and Restaurant (RE). The age of acquisition mastery scale is in number of months.

Similarly to [29], gestural effort, which is the amount of articulatory change within an utterance, was examined for the ten read speech protocols. The gestural measure was used to analysed seventeen different Chomsky-Halle articulatory manners using binary representations, whereby the greater number of switches between phoneme manners resulted in a higher gestural effort value (i.e., Hamming distance). It is believed that more rapid manner activation/non-activation productions between each phoneme necessitates increased fine motor control (e.g., motoric coordination). Moreover, [29] found that recorded utterances with a higher gestural effort measure produced improved automatic depression detection. In Figure 4, gestural effort analysis demonstrated that 'Arthur' was the broadest in terms of individual sentence ranges. Based on mean gestural effort, the 'John' and 'Restaurant' read speech protocols were the most demanding.

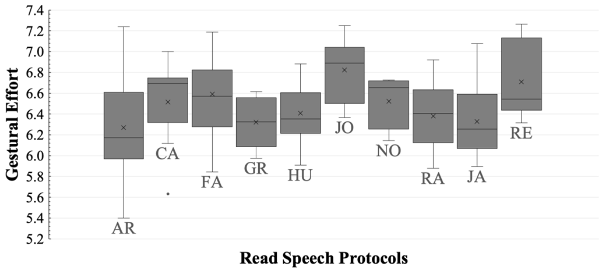

Figure 4: Sentence-level distributions of gestural effort per read speech protocol: Arthur (AR), Caterpillar (CA), Farmer (FA), Grandfather (GR), Hunter (HU), John (JO), Northwind (NO), Rainbow (RA), Jazz (JA), and Restaurant (RE).

It is mostly unknown what kind of impact multi-dimensional criteria aspects have on automatic speech-based illness detection. But future investigation of speech-based digital health screening applications using read speech protocol may reveal improved screening capabilities, especially new protocols tailored towards specific illnesses/disorders. While it is known that criteria (e.g., phonetic, linguistic, affective) influence each other and the acoustic speech signal, there is no known protocol that simultaneously covers the entire ranges of these criteria. Figure 5 shows that even within considering three phoneme-based criteria, read speech protocol ranges can vary. For example, in Figure 5, it is observed that the 'Arthur' protocol is more suitable for younger test subjects, wherein a longer and a more repetitive token word sample is required (i.e., but it has unsuitable death theme). On the contrary, the 'Restaurant' protocol is better suited for older test subjects, wherein a shorter and a less repetitive token word sample is desirable. Knowledge regarding criteria restraints within read speech protocols further provides better insight towards which specific read speech protocol to select for health screening purposes.

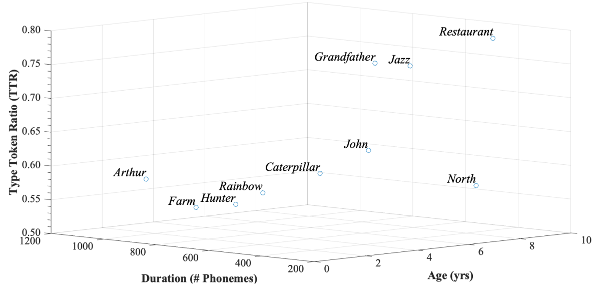

Figure 5: Multidimensional read speech protocol criteria. Depending on speaker demographic and digital health criteria needs, careful consideration should be taken to make sure protocols are appropriate for digital health applications.

2.3. Linguistic and Affective Criteria

The 'Farm', 'John', 'North Wind', 'Rainbow', and 'Jazz' read speech protocols include some sentences that contained passive voice grammar. However, the passive voice maximum for all protocols examined was below 40%, indicating that new protocols could investigate greater use of passive voice. For example, it is well-known that individuals with dementia usually exhibit poor cognitive function, word retrieval difficulty, increased speech disfluencies, and struggle with constructing or recalling novel appropriate responses, especially from passive voice texts [36].

The reading level for the read ten speech protocols were calculated using the Flesch-Kincaid method for average grade level and average reading ease [31]. Analysis in Table 2 indicated that the 'Restaurant' (9.5) and 'North Wind' (8.2) read speech protocols may be inappropriate for evaluating speech of individuals younger than eighth grade based on reading ease scores. Strangely, the reading ease score for the 'Rainbow Passage' (7.7) somewhat contradicts [9], where the introduction suggests its broad use for '*school-aged children*'.

Because reading is a requirement for these protocols, linguistic norm criteria should be closely examined to establish age appropriateness and/or level of difficulty [3, 31, 32]. In Table 2, the average age of word exposure for 'Jazz' was the highest (6.4) due to many less frequent keywords. The 'Jazz' and 'Restaurant' read protocols also demonstrated the highest values for average word acquisition (6.58, 6.30), whereas 'Farm' contained average word acquisition that was much lower (4.86). Table 2 analysis demonstrates that word grade level, reading ease, exposure, and acquisition are different criteria that do not always directly associate with each other.

In Table 2, using affective norms text-processing software [33], an examination of affect across the ten different read speech protocols reveals that the degree of arousal is narrow and typically in the neutral range. Dominance was also shown to be typically neutral for the read speech protocols, except for the 'North Wind', which had greater dominance (6.40). Moreover, valence average per read speech protocol demonstrated a bias towards positive valence keywords. The only exception to this was the 'Arthur' read speech protocol that had a neutral valence value of 4.93, nearing the negative valence range (≤ 4.50). Affective results herein show that there is a need for read speech protocols where arousal, dominance, and valence are more extreme to understand paralinguistic behaviours in different emotional contexts. A recent speech-based study [30] on automatic mood disorder detection indicated that protocols containing a broader range of valence can help improve detection of individuals with mental illness.

3. Conclusion

It is vital that a greater number of criteria are taken into account than most current speech-based studies when designing, choosing, and executing speech elicitation materials for digital health analysis. It is likely that deliberately tailored protocol design for speech-based digital health applications will produce greater benefit, with less computation delving through unnecessary excess acoustic speech data to analyze only a small percentage. Also, protocol designs may have advantages over free speech collections because they can potentially isolate speech and/or language tasks that might otherwise involve a mixture of cognitive, memory, and motor articulation skill level. Some isolated read verbal tasks already exist, such as the diadochokinetic and Stroop color test protocols, but their speech focus is more obvious to the participant, repetitive, and less natural – which may influence speech behaviors.

Generally, there is much greater room for new development in the field of speech-language elicitation protocols for health screening purposes. Future studies on illness/disorders should directly measure many different read speech protocol sensitivity/specificity results to help determine which are the best depending on the illness of interest and test subject age.

Table of Contents for this manuscript

4. References

[1] Chevrie-Muller, C., Sevestre, P., & Seguier, N., “Speech and psychopathology”, Lang. and Speech, vol. 28 (1), pp. 57-79, 1985.

[2] Hirschberg, J., Hjalmarsson, A., & Elhadad, N., “You’re as sick as you sound: using computational approaches for modeling speaker state to gauge illness and recovery”, In: A. Neustein (ed.) Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, Springer Science + Business Media, pp. 305-322, 2010.

[3] MacMahon, M.K.C., “The woman behind ‘Arthur’”, J. Intern. Phonetic Assoc., vol. 21 (1), pp. 29-31, 1991.

[4] Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E., & Russell, S., “The caterpillar: a novel reading passage for assessment of motor speech disorders”, Am. J. Spch. Lang. Path., vol. 22 (1), pp. 1-9, 2013.

[5] Crystal, T.H., & House, A.S., “Segmental duration in connected speech signals: preliminary results”, J. Acoust. Soc. Am., vol. 72 (3), pp. 705-716, 1982.

[6] Van Riper, C., Speech Correction (4th Ed.), Prentice Hall, Englewood Cliffs, NJ – USA, 1963.

[7] Tjaden, K., & Wilding, G., “Rate and loudness manipulations in dysarthria: acoustic and perceptual findings”, J. Speech Lang. Hear. Res., vol. 47 (4), pp. 766-783, 2004.

[8] Townsend, G.F., Aesop’s Fables, George Routledge & Sons, London & New York, 1868.

[9] Fairbanks, G., Voice and Articulation Drillbook (2nd Ed.), Harper & Row, New York – USA, pp. 124-139, 1960.

[10] Egan, J.P., “Articulation testing methods”, Laryngoscope, vol. 58, pp. 955-991, 1948.

[11] Patel, R.R., Awan, S.N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Svec, J., and Hillman, R., “Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function”, American J. Speech-Language Pathology, vol. 27 (3), pp. 887-905, 2018.

[12] Schiel, F., Draxler, C., Baumann, A., Ellbogen, T., & Steffen, A., “The production of speech corpora”, Version 2.5, Bavarian Arch. for Speech Signals, University of Munich, 2004.

[13] Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., and Geralts, D.S., “Voice acoustic measure of depression severity and treatment response collected via interactive voice response (IVR) technology”, J. Neurolinguistics, vol. 20 (1), pp. 50-64, 2011.

[14] Jaing, H., Hu, B., Liu, Z., Lihua, Y., Wang, T., Liu, F., Kang, H., and Li, X., “Investigation of different speech types and emotions for detecting depression using different classifiers, Speech Comm., vol. 90, pp. 39-46, 2017.

[15] Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B., and Mehta, D.D., “Vocal biomarkers of depression based on motor incoordination”, In: Proc. AVEC '13: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 41-48, 2013.

[16] Ludlow, C.L., Kent, R.D., and Gray, L.C., Measuring Voice, Speech, and Swallowing in the clinic and Laboratory, Plural Publishing Inc., San Diego, CA – USA, 2019.

[17] Lammert, A.C., Melot, J., Sturim, D.E., Hannon, D.J., DeLaura, R., Williamson, J.R., Ciccarelli, G., & Quatieri, T.F., “Analysis of phonetic balance in standard English passages”, J. Speech, Lang., and Hearing Res., vol. 63 (4), pp. 917-930, 2020.

[18] Powell, T.W., “A comparison of English reading passages for elicitation of speech samples from clinical populations”, Clinical Linguistic & Phonetics, vol. 20, pp. 91-97, 2006.

[19] Kent, R.D., Kent, J.F., & Rosenbek, J.C., “Maximum performance tests of speech production”, J. of Speech Hear. Disord., vol. 52, pp. 367-387, 1987.

[20] Deterding, D., “The north wind versus a wolf: short text for the description and measurement of English pronunciation”, J. Intern. Phonetic Assoc., vol. 36 (2), pp. 187-196, 2006.

[21] Bayestehtashk, A., Asgari, M., Shafran, I., and McNames, J., “Fully automated assessment of the severity of Parkinson’s disease from speech”, Comp. Speech Lang., vol. 29 (1), pp. 172-185, 2015.

[22] Perez, M., Jin, W., Le, D., Carlozzi, N., Dayalu, P., Roberts, A., and Mower-Provost, E., “Classification of Huntington disease using acoustic and lexical features”, In: Proc. INTERSPEECH 2018, pp. 1898-1902, 2018.

[23] Turner, G.S., Tjaden, K., and Weismer, G., “The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis”, J. Speech, Lang., and Hearing Res., vol. 38 (5), pp. 1001-1013, 1995.

[24] Whitfield, J.A., Kriegel, Z., FullenKamp, A.M., and Mehta, D.D., “Effects of concurrent manual task performance on connected speech acoustics in individuals with Parkinson disease”, J. Speech, Lang., Hearing Res., vol. 62 (7), pp. 2099-2117, 2019.

[25] Howell, P., Sackin, S., and Glenn, K., “Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word segment markers”, J. Speech, Lang., Hearing Res., vol. 40, pp. 1085-1096, 1997.

[26] Yunusova, Y., Weismer, G., Kent, R.D., and Rusche, N.M., “Breath-group intelligibility in dysarthria: characteristics and underlying correlates”, J. Speech, Lang., and Hearing Res., vol. 48, pp. 1294-1310, 2005.

[27] Ben-David, B.M., Moral, M., Namasivayam, A., & van Lieshout, P., “Linguistic and emotional-valence characteristics of reading passages for clinical use and research”, J. of Fluency Disord., vol. 49, pp. 1-12, 2016.

[28] Stasak, B., Epps, J., & Goecke, R., “Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect”, INTERSPEECH ’17, Stockholm – Sweden, pp. 834-838, 2017.

[29] Stasak, B., Epps, J., and Lawson, A., “Analysis of phonetic markedness and gestural effort measures for acoustic speech-based depression classification”, In: Proc. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), San Antonio, TX – USA, pp. 165-170, 2017.

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

[31] Kincaid, J.P., Fishburne, R.P., Rogers, R.L., and Chissom, B.S., “Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel”, U.S. Naval Research Branch Report, pp. 8-75.

[32] Kyle, K. & Crossley, S.A., “Automatically assessing lexical sophistication: Indices, tools, findings, and application”, TESOL Quarterly, vol. 49 (4), pp. 757-786, 2015.

[33] Crossley, S.A., Kyle, K., & McNamara, D.S., “Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis”, Behavior Research Methods, vol. 49 (3), pp. 803-821, 2017.

[34] Carnegie Mellon University (CMU), 1993. The Carnegie Mellon Pronouncing Dictionary v0.1. Carnegie Mellon University: http://www.speech.cs.cmu.edu/cgi- bin/cmudict

[35] Mines, M.A., Hanson, B.F., and Shoup, J.E., “Frequency of occurrence of phonemes in conversational English”, Language and Speech, vol. 21 (3), pp. 221-241, 1978.

[36] Emery, V.O.B., “Language impairment in dementia of the Alzheimer type: a hierarchical decline?”, Inter. J. Psych. Medicine, vol. 30 (2), pp. 145-164, 2000.